

Department of STATISTICS AND QUANTITATIVE METHODS

PhD program: STATISTICS AND MATHEMATICS FOR FINANCE

Cycle: XXXI

Curriculum in: MATHEMATICAL FINANCE

FAST MASS COMPUTATION OF SENSITIVITIES AND EFFECTIVE HEDGING OF FINANCIAL PRODUCTS

Surname: DALUIISO

Name: ROBERTO

Registration number: 802676

Tutor: PROF. FABIO BELLINI

Supervisor: PROF. MASSIMO MORINI

Coordinator: PROF. GIORGIO VITTADINI

ACADEMIC YEAR 2017-2018

Contents

Introduction	v
Notation	ix
I Hedging with multiple sensitivities	1
1 Hedging under correlation	3
1.1 Introduction	3
1.2 Hedging efficiently in illiquid markets	5
1.3 Realistic optimal hedging	9
1.4 Practical case study	14
1.5 Conclusions	22
2 Hedging under recalibration	23
2.1 Introduction	23
2.2 Formalization of the problem	25
2.3 Indetermination	31
2.4 A simple case	34
2.5 The general case	35
2.6 Linearity of Greeks and hedged portfolios	37
2.7 Comparison to the derivatives of bestfit calibration	39
2.8 Considerations on the choice of the representative	42
2.9 Numerical experiments	43
2.10 Conclusions and further research	47
II Fast computation of multiple sensitivities	51
3 Fast first order sensitivities for discontinuous payoffs	53
3.1 Introduction	53
3.2 The digital option	55
3.3 Arbitrary combinations of indicator functions	56
3.4 Algorithmic differentiation and preprocessing	58
3.5 Monte Carlo	62
3.6 Numerical examples	67
3.7 Conclusions	80

4	Second order sensitivities in linear or constant time	83
4.1	Introduction	83
4.2	Setting	85
4.3	Review of existing algorithms	86
4.4	New methods	91
4.5	Empirical results	95
4.6	Conclusions	105
A	Technical appendix	109
A.1	Hedging with any number of correlated assets	109
A.2	First order distributional differentiation	110
A.3	Call spread analytical sensitivities	115
A.4	Second order distributional differentiation	118
	Bibliography	121
	Acknowledgements	125

Introduction

In the context of mathematical finance, sensitivities are derivatives of a price with respect to a risk factor or to a parameter of a pricing model. The present thesis is dedicated to the development of a toolkit for their computation and practical use in multi-dimensional settings. Specifically, we want to address two main points, corresponding to the two parts of this work:

1. The traditional theoretical setting where sensitivities-based hedging is justified involves questionable idealizations, such as continuous-time portfolio rebalancing with no costs. Do more realistic assumptions impact the way in which sensitivities should be used?
2. When the number of drivers is very large, the estimation of sensitivities becomes a computationally demanding task. How can many of them be calculated efficiently?

Sensitivities or Greeks have played a crucial role in the foundations of modern option pricing, being at the heart of the original Delta-based replication argument in Black and Scholes (1973). Theoretical developments in the last decades have correctly pointed to the many oversimplifications in that pioneering approach, which requires continuous frictionless trading. This literature often throws sensitivities out of the picture of derivatives risk management, in favour of more complex hedging strategies relying for instance on stochastic control. In many markets, practice has not followed, and traders base most of their portfolio immunization activity on sensitivities. Why is this the case?

The reasons are probably manifold, but we believe that one of them is scalability. When one has to hedge at the same time multiple sources of risk, complex optimizations become quickly unfeasible; moreover, they are very sensitive to the multivariate distributional assumptions of the dynamical model used in pricing, which is most often the result of compromises between realism and parsimoniousness/estimability, and which does not even stay fixed, since its parameters are frequently recalibrated to new market data. In contrast, the elementary recipe which prescribes to keep the sensitivities close to zero has an essentially model-free interpretation, and in its simplicity, it is computationally less subject to the curse of dimensionality.

Traders working in these model-risky and highly-dimensional settings, where exact optimality results are of little applicability, have almost been left alone by academicians. This thesis aims at showing that mathematics has something useful to suggest even if one takes the fact that the sensitivities must be the main input of decision as a given because of all the above concerns.

Overview of Part I

Part I is concerned with the effective use of multiple sensitivities in practice.

Chapter 1 studies the effects on hedging of the interaction between different underlying instruments as modelled by instantaneous diffusive correlation. This parameter would not play any role in idealized continuously-rebalanced hedging, but we find that it does if rebalancing times are in finite number and potentially different for different instruments, as often in practice. Under suitable assumptions, we find a strategy in which the sensitivities are combined in a nontrivial way, since some hedge positions are sometimes not rebalanced because the corresponding exposure can be in part offset by overweighting or underweighting other correlated hedges.

This chapter, which is joint work with prof. Massimo Morini, has appeared with minor differences on the journal “Quantitative Finance” in 2017 (Daluiso and Morini, 2017).

Chapter 2 considers how the practice of periodically recalibrating model parameters to market data affects the way in which sensitivities should be looked at. The fact that these parameter changes cannot be ignored is reflected into another widespread practice, namely that of monitoring sensitivities to model parameters. However, recalibration effectively falsifies the distributional assumptions behind the pricing model, so that a formalization is almost hopeless inside a traditional stochastic processes based no-arbitrage theory. Hence we propose an alternative mathematisation based on differential geometry, which describes the degrees of freedom one has in the construction of the hedging portfolio in this setting.

This chapter was presented at the 9th World Congress of the Bachelier Finance Society in New York in July 2016, and at the XVIII Quantitative Finance Workshop in Milan in January 2017.

Overview of Part II

Part II focuses on the efficient computation of large numbers of sensitivities.

Chapter 3 concentrates on first order sensitivities of prices whose computation is costly due to the need of Monte Carlo simulation. Our starting point is that for continuous payoffs, the pathwise application of a computer science technique known as adjoint algorithmic differentiation gives remarkably fast and accurate price gradients of arbitrary length; however, the generalizations to discontinuous payoffs like digital options are nontrivial. The new algorithm proposed here distinguishes itself by extending the pathwise adjoints method in a most natural way, and by its empirically very low Monte Carlo uncertainties.

The results of this chapter have been the subject of a talk both at the 2nd International Conference on Computational Finance in Lisbon in September 2017, and in the XIX Quantitative Finance Workshop in Rome in January 2018. The present text, with minor modifications, has been published on the “International Journal of Theoretical and Applied Finance” in 2018, coauthored by dr. Giorgio Facchinetti (Daluiso and Facchinetti, 2018).

Chapter 4 looks for fast algorithms to compute the full second order sensitivity matrix of a Monte Carlo price. Many combinations of first order estimators have been tried in the literature to this purpose, and our first contribution is an orderly theoretical and empirical comparison of these proposals. Then, since none of the alternatives appears satisfactory in all settings, we propose two original methods: the first one generalizes the idea of the previous chapter, while the other one leverages a functional

relation between first and second order derivatives. The former shows excellent generality and computational times. The latter has more limited applicability, but it is by far the most effective in at least one relevant example, and has a theoretical interest, being the first practical estimator of the full Hessian whose complexity, as a multiple of that of the only-price implementation, does not grow with the dimension of the problem.

These findings have been presented at the 10th World Congress of the Bachelier Finance Society in Dublin in July 2018.

Notation

Throughout this thesis, the following notational conventions are adopted. Symbols not listed here should be defined in the chapter where they are used.

Scalars and vectors

- Scalars are typeset in italic (e.g. a).
- Vectors are typeset in boldface italic (e.g. $\mathbf{a}$ has components a_i), and are interpreted as columns unless transposed (e.g. $\mathbf{a}$ is a column and $\mathbf{a}^\top$ is a row). In particular, the gradient $\nabla P = \text{grad } P$ of a scalar function P is a column vector.
- Fixed a target scalar function P , the adjoint of a column vector $\mathbf{a}$ is $\bar{\mathbf{a}} = \partial P / \partial \mathbf{a}$ and is interpreted as a row vector.
- The concatenation $(\mathbf{a}, \mathbf{b})$ of two column vectors is a column vector, while the concatenation $(\bar{\mathbf{a}}, \bar{\mathbf{b}})$ of two row vectors is a row vector.
- The scalar product of two vectors $\mathbf{a}$ and $\mathbf{b}$ is denoted by $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^\top \mathbf{b}$, while the dot symbol $\cdot$ will always denote standard (matrix) multiplication without any transposition, and is often omitted.
- Given a vector $\mathbf{v} \in \mathbb{R}^d$ and $k \in \{1, \dots, d\}$, we denote by $\mathbf{v}_{-k} \in \mathbb{R}^{d-1}$ the vector obtained from $\mathbf{v}$ removing its k -th component, and with $\mathbf{v}(v_k = x)$ the vector obtained from $\mathbf{v}$ substituting the k -th component with the value x . Moreover, for a function $\psi : \{0, 1\}^h \rightarrow \mathbb{R}$ and $i = 1, \dots, h$, we define $\Delta^{(i)}\psi(\mathbf{a})$ as the difference $\psi(\mathbf{a}(a_i = 1)) - \psi(\mathbf{a}(a_i = 0))$.

Matrices

- Matrices are typeset in straight boldface (e.g. $\mathbf{A}$ has entries A_{ij}).
- The Jacobian of a vector-valued function $\mathbf{f}$ is denoted by $D\mathbf{f}$, or by $D_{\boldsymbol{\theta}}\mathbf{f}$ to specify the set of variables $\boldsymbol{\theta}$ with respect to which differentiation is performed; it is obtained by stacking the transposed gradients of the function components f_i . The Hessian of a scalar function P is $\text{Hess } P$ or $\text{Hess}_{\boldsymbol{\theta}} P$ and is defined as the Jacobian of $\text{grad } P$ (resp. $\text{grad}_{\boldsymbol{\theta}} P$).
- Fixed a target scalar function P , the adjoint of a matrix $\mathbf{A}$ is the matrix $\bar{\mathbf{A}} = \partial P / \partial \mathbf{A}$ with components $\bar{A}_{ij} = \partial P / \partial A_{ji}$ (note the inversion of indices).
- The identity matrix of dimension n is denoted by $\mathbf{I}_n$, while $\mathbf{0}$ denotes a matrix or vector full of zeros.
- If $\mathbf{a}$ is a vector, then $\text{diag}(\mathbf{a})$ is the diagonal matrix whose diagonal is $\mathbf{a}$.

Probability

- Unless otherwise stated, all probabilistic statements refer to a fixed probability space, whose probability measure is denoted by $\mathbb{P}$. Expectation is denoted by $\mathbb{E}$. Sometimes we will endow this space with a filtration $(\mathcal{F}_t)_{t \geq 0}$, and will denote by $\mathbb{E}_t$ the conditional expectation operator with respect to $\mathcal{F}_t$.
- The variance of a random variable X is denoted by $\text{Var}(X)$; its covariance with another random variable Y is denoted by $\text{Covar}(X, Y)$.
- I_E is the indicator function of a set E . If $\mathbf{v}$ is a random vector, then $\mathbf{I}_{\mathbf{v} > \mathbf{0}}$ denotes the vector $(I_{\{v_1 > 0\}}, \dots, I_{\{v_d > 0\}})$.
- $\mathcal{N}(\mathbf{m}, \Sigma)$ is the Gaussian distribution with mean vector $\mathbf{m}$ and covariance matrix Σ ; $\mathcal{N}(\cdot)$ also denotes the cumulative distribution function of the univariate standard normal $\mathcal{N}(0, 1)$.
- For stochastic processes, we write the time argument in the subscript (e.g. $\mathbf{X}_t$) if no confusion arises, or as a function argument (e.g. $X_i(t)$) otherwise.
- If X_t and Y_t are stochastic processes, then $\langle X, Y \rangle_t$ denotes their quadratic covariation process.

Geometry

- If f is a function and E is a set, then $f|_E$ is the restriction of f to the set E ; if f is injective, then f^{-1} denotes its inverse (not its reciprocal $1/f$). If f and g are functions, then $f \circ g$ is their composition.
- The differential of a smooth function f is denoted by df . Its value at the point $\mathbf{y}$ is denoted by $d_{\mathbf{y}}f$.
- $\text{span}(\mathbf{a}_i)_{i \in I}$ is the linear space generated by a set of vectors $(\mathbf{a}_i)_{i \in I}$.
- $A^\perp$ denotes the set of vectors which are orthogonal to all elements in the set A .
- If M is a manifold and $\mathbf{m} \in M$, then $T_{\mathbf{m}}M$ is the tangent space to M in point $\mathbf{m}$, and $T_{\mathbf{m}}^*M$ is its dual (the cotangent space).

Part I

Hedging with multiple sensitivities

Chapter 1

Hedging under correlation

In this chapter, we show that when a derivative portfolio has different correlated underlyings, hedging using classical Greeks (first-order derivatives) is not the best possible choice. We first show how to adjust Greeks to take correlation into account and reduce P&L volatility. Then we embed correlation-adjusted Greeks in a global hedging strategy that reduces cost of hedging without increasing P&L volatility, by optimization of hedge re-adjustments. The strategy is justified in terms of a balance between transaction costs and risk-aversion, but, unlike more complex proposals from previous literature, it is completely defined by observable parameters, geometrically intuitive, and easy to implement for an arbitrary number of risk factors. We test our findings on a CVA hedging example. We first consider daily re-hedging: in this test correlation-adjusted Greeks allow to reduce P&L volatility by more than 30% compared to standard Deltas. Then we apply our general strategy to a context where a CVA portfolio is exposed to both credit and interest rate risk. The strategy keeps P&L volatility in line with daily standard Delta-hedging, but with massive cost-saving: only six rebalances of the illiquid credit hedge are performed, over a period of six months.¹

1.1 Introduction

Traders are often required to hedge the risks of portfolios that depend on different correlated risk factors. XVA (Credit, Funding and Capital Value Adjustments) hedging, hybrid portfolios, and Delta-Vega hedging with a stochastic volatility model are all examples of the issue. It is claimed by market wisdom that such hybrid exposures can be efficiently hedged only by strategies that take explicitly into account the correlation between the different risk factors, otherwise “cross-gammas” will eat out the trader’s profit. Yet there is no standard recipe on how this should be done. In this work we clarify the issue, by showing under which conditions a hedging strategy based on correlations can reduce P&L volatility, and how such a strategy can be designed.

Curiously, up to the present day the practice of hedging in financial markets still draws its theoretical foundations from the Black and Scholes original approach, that assumes continuous hedging and perfect immunization from all risks. Yet, hedging in real markets is dramatically at odds with this theoretical framework. Some of the risk factors driving the value of a derivative are not tradable or they are illiquid, and in any

¹A slightly different version of this chapter has been published in Daluio and Morini (2017).

case none of them is so liquid to be traded continuously. Additionally, traders tend to rebalance their hedges at different frequencies. For example Delta-hedging, usually accomplished with basic and liquid linear assets, is performed at higher frequency than Vega-hedging which is based on buying and selling less liquid options. Similarly, in hedging CVA (Credit Value Adjustment) or rate-credit hybrids, the rebalancement of the credit component in the hedging portfolio, usually represented by rather illiquid CDS, cannot be performed with the same frequency that is used in rebalancing the rate part of the hedging portfolio.

In this work we show that when some risk factors are not tradable, or are traded and hedged at lower frequency compared to other hedges, the Black and Scholes hedging recipe based on Greeks (first order derivatives of the price function) is suboptimal and P&L volatility can be minimized only using a strategy that takes correlations into account. The problem arises from the presence of transaction costs and from the risk aversion of traders, but the solution we find does not require knowledge of the level of transaction costs nor knowledge of the level of the risk aversion of the trader. This is a plus of our approach since such levels are difficult to estimate. The only input we need is a piece of information usually available in market reality: the frequency of rebalancement of the different components of the hedging portfolio, associated to the trader's tolerance to P&L volatility.

A rich and interesting literature exists on the topic of imperfect pricing and hedging taking transaction costs into consideration, see for example Hodges and Neuberger (1989); Whalley and Wilmott (1997); Zakamouline (2005), but it never really impacted the everyday management of derivatives. One of the reasons for that is most of the literature concentrating on linear transaction costs, while real transaction costs have strong non-linearities, that are central in the present analysis. The literature on adjusting Deltas based on the correlation between local/stochastic volatility parameters and the underlying had some more impact on market practice, and it is wide and relevant to the work presented here. For example Crépey (2004) explores, for an option on a single underlying, how dependence between changes in the underlying and changes in implied volatilities may affect the performance of two different hedging strategies in discrete time, while Alexander and Nogueira (2004) compute, for a general class of local/stochastic volatility models, an adjustment factor for Delta, Gamma and Theta that takes correlation with instantaneous volatility into account and improves empirical hedging performance. In Bartlett (2006) similar adjustments are computed for the case of the SABR model, while Mercurio and Morini (2009) take a heuristic approach to see how traders adjust hedges to overwrite the Deltas coming from local and stochastic volatility models. Our work starts from these contributions, but generalizes them under several respects. We move from the issue of improving Delta-hedging using equity-volatility correlation to the general setting of hedging a portfolio of products depending on many correlated assets. CVA, volatility and hybrid trading are all possible applications. In this context, we extend from the case of hedging with one single underlying, covered in the above literature on Delta-hedging, to the case of hedging with more correlated risk factors. We show that also in this case there are corrections, depending on correlation, to model-based hedges, that can improve hedging performance. Finally, we link the two streams of literature above, giving a general strategy where the decision to leave some factors unhedged, and the related need to correct model hedges based on correlation, is taken based on the trade-off between trader's tolerance to PnL volatility and transaction costs.

The rest of the work is organized as follows. In Section 1.2 we show how an optimal hedge must be built under the assumption that the different rebalancing times of the

different components of the hedging portfolio have been exogenously determined by the trader, starting with a situation where hedging is performed on one single risk factor. This usually happens when the trader thinks it is not cost-effective to hedge the second risk-factor. Next we consider what to do when an increased exposure to the second risk factor makes it necessary to hedge also the second asset. Then in Section 1.3 we see how these optimal hedges can be embedded in a consistent and rational hedging strategy, taking into account hedging costs and risk aversion but building a strategy that does not require any assumption on unobservable variables. Both financial and geometric intuition on the soundness of the strategy is given. Finally in Section 1.4 we test our results on a numerical example of CVA hedging. Appendix A.1 covers the extension to a generic number of underlying risks.

1.2 Hedging efficiently in illiquid markets

We have a derivative whose price depends on two assets F_t and C_t , that represent both the factors driving the risk of the derivative, and the assets that one buys and sells to hedge these risks. For example the derivative could be a rates-credit hybrid so that F_t is associated to interest rate risk and C_t is associated to credit risk, or F_t could be the underlying of an option-like derivative, bought or sold to hedge the Delta, while C_t could be a vanilla option used to hedge a Vega exposure. The price is given by a model formula

$$\Pi_t = \Pi(F_t, C_t).$$

and the two assets have diffusive dynamics

$$\begin{aligned} dF_t &= \mu_F(t) dt + \sigma_F(t) dW_t^F, \\ dC_t &= \mu_C(t) dt + \sigma_C(t) dW_t^C, \end{aligned}$$

where by dW^x we indicate the stochastic drivers, all equally distributed. We assume they are correlated with correlation ρ . This means that the shock dW_t^C on C can be written as

$$dW^C = \sqrt{1 - \rho^2} dW^{C\perp} + \rho dW^F,$$

where $dW^{C\perp}$ is the idiosyncratic C shock independent of dW^F . This way the conditional distribution of dW^C is

$$(dW^C | dW^F = X) \sim \rho X + \sqrt{1 - \rho^2} dW^{C\perp}.$$

The processes μ_X and σ_X will be supposed to be continuous in time. Their concrete form will not play any role in the subsequent derivation. We will often drop the dependency on t in the notation, writing for instance σ_X instead of $\sigma_X(t)$.

We will suppose that both the derivative and the hedging instrument do not provide payments in the time span of interest. Otherwise, one should imagine that each of Π , F , C represents the value of a self-financing trading strategy reinvesting the cash proceedings in a locally risk-free account.

1.2.1 First hedge: Rates (or Delta)

This subsection restates some results that can be found, with some differences and in different contexts, in Crépey (2004); Bartlett (2006); Mercurio and Morini (2009); Alexander and Nogueira (2004). In the next subsection we extend this to a second hedge. Appendix A.1 generalizes to n hedges.

Heuristic derivation

In some cases one of the risk factors above is not hedged at inception, because exposure is not yet sufficient to justify the cost of buying a hedge in a non-perfectly-liquid market. Suppose this is the situation of the factor C_t . On the other hand, either the exposure to F_t risk factor is higher or the associated fixed transaction costs are lower, so that the trader considers convenient to invest in F_t as a hedge. Thanks to correlation between the two assets, a trader can hedge also part of its exposure to credit/volatility C even with a hedging portfolio solely based on rates/underlying F . We start from the fact that

$$\mathbb{E} [dW^C dW^F] = \rho dt \quad \Rightarrow \quad \mathbb{E} [dW^C | dW^F] = \rho dW^F,$$

and, following Morini (2011), we build first some intuition in the driftless case $\mu_F = \mu_C = 0$, working with discrete changes of the underlying factors like those observed in the market. In hedging we want to know how much of F we have to buy today to compensate the change in the value of the derivative in case F moves overnight by some discrete amount ΔF . Thus we are going to estimate

$$\frac{\Delta \Pi(F_t, C_t)}{\Delta F} = \frac{\Pi(F_{\Delta t}, C_{\Delta t}) - \Pi(F_0, C_0)}{\Delta F}.$$

Assuming a shock ΔF over one day corresponds to assuming that we are working with a shock of the stochastic driver over one day given by

$$\Delta W^F \approx \frac{\Delta F}{\sigma_F} =: \bar{\Delta}.$$

This corresponds to an expected shock of the C driver

$$\mathbb{E} [\Delta W^C | \Delta W^F = \bar{\Delta}] = \rho \bar{\Delta},$$

leading to an expected shock of the C asset

$$\mathbb{E} [\Delta C_t | \Delta F] \approx \sigma_C \mathbb{E} [\Delta W^C | \Delta W^F = \bar{\Delta}] = \sigma_C \rho \bar{\Delta}.$$

Thus to hedge the expected movement of Π we can sell the following amount of the F asset:

$$\begin{aligned} & \frac{\Pi(F_0 + \Delta F, C_0 + E[\Delta C | \Delta F]) - \Pi(F_0, C_0)}{\Delta F} \\ & \approx \frac{\Pi(F_0 + \sigma_F \bar{\Delta}, C_0 + \sigma_C \rho \bar{\Delta}) - \Pi(F_0, C_0)}{\sigma_F \bar{\Delta}} \end{aligned} \quad (1.1)$$

Formal derivations as variance minimization

When the unknown quantity $\bar{\Delta}$ goes to zero as $\Delta t \rightarrow 0$, the hedge computed in (1.1) converges to the following quantity of F asset:

$$g_F = \frac{\partial \Pi}{\partial F} + \rho \frac{\sigma_C}{\sigma_F} \frac{\partial \Pi}{\partial C}, \quad (1.2)$$

as is clear from Taylor expansion of the numerator as a function of $\bar{\Delta}$.

The same result can be alternatively obtained as the quantity g_F of the asset F we need to hold if we want $\Pi_{\Delta t} - g_F F_{\Delta t}$ to have the smallest possible variance, in the

limit $\Delta t \rightarrow 0$. Only for this argument, we will use the expressive notations dC , dF , dt to denote ΔC , ΔF , Δt in the limit $\Delta t \rightarrow 0$.

Since

$$d\Pi_t \approx \frac{\partial \Pi}{\partial F} dF + \frac{\partial \Pi}{\partial C} dC + \dots dt,$$

we are essentially trying to minimize

$$\text{Var} \left[\left(\frac{\partial \Pi}{\partial F} - g_F \right) dF + \frac{\partial \Pi}{\partial C} dC \right] := \text{Var} \left[a dF + \frac{\partial \Pi}{\partial C} dC \right].$$

Classical considerations tell us that the minimizer is $a = -\frac{\text{Covar}(dC, dF)}{\text{Var}(dF)} \frac{\partial \Pi}{\partial C}$, i.e.

$$g_F - \frac{\partial \Pi}{\partial F} = \frac{\text{Covar}(dC, dF)}{\text{Var}(dF)} \frac{\partial \Pi}{\partial C} = \frac{\sigma_C \sigma_F}{\sigma_F^2} \rho \frac{\partial \Pi}{\partial C} = \boxed{\frac{\sigma_C}{\sigma_F} \rho \frac{\partial \Pi}{\partial C}}.$$

This local minimization of variance yields the same result as a minimization of the total risk-adjusted variance of the discounted payout from inception to maturity, as we see in the following. We want to minimize $\text{Var}[\tilde{\Pi}_T - \tilde{H}_T]$, where H_t is the value at time t of the hedging portfolio and a tilde indicates a discounted value. By self-financing

$$\tilde{H}_T = H_0 + \int_0^T g_F d\tilde{F}_t$$

and by classical Delta-hedging

$$\tilde{\Pi}_t = \Pi_0 + \int_0^T \frac{\partial \Pi}{\partial F} d\tilde{F}_t + \int_0^T \frac{\partial \Pi}{\partial C} d\tilde{C}_t, \quad (1.3)$$

hence we have to choose $a := \left(\frac{\partial \Pi}{\partial F} - g_F \right)$ minimizing:

$$\begin{aligned} \text{Var} \left[\int_0^T \left(\frac{\partial \Pi}{\partial F} - g_F \right) d\tilde{F}_t + \int_0^T \frac{\partial \Pi}{\partial C} d\tilde{C}_t \right] = \\ \int_0^T \mathbb{E} \left[D(0, t)^2 \left(a^2 \sigma_F^2 + \left(\frac{\partial \Pi}{\partial C} \right)^2 \sigma_C^2 + 2\rho \sigma_C \sigma_F a \frac{\partial \Pi}{\partial C} \right) \right] dt. \end{aligned}$$

where $D(0, t)$ is the discount factor. The integrand can be pointwise minimized, choosing

$$a = -\rho \frac{\sigma_C}{\sigma_F} \frac{\partial \Pi}{\partial C}, \quad \text{i.e.} \quad g_F = \frac{\partial \Pi}{\partial F} + \rho \frac{\sigma_C}{\sigma_F} \frac{\partial \Pi}{\partial C}.$$

Note that the strategy does not depend on the horizon T . It also minimizes pathwise the quadratic variation

$$\int_0^T D(0, t)^2 \left(a^2 \sigma_F^2 + \left(\frac{\partial \Pi}{\partial C} \right)^2 \sigma_C^2 + 2\rho \sigma_C \sigma_F a \frac{\partial \Pi}{\partial C} \right) dt,$$

which is a proxy of what traders call *Profit&Loss volatility*, i.e. the sum of squared wealth movements, which is usually taken as a measure of hedging effectiveness. P&L volatility tends to our quadratic variation when the frequency of monitoring tends to zero and neglecting the discounting terms.

1.2.2 Second hedge: Credit (or Vega)

At some point, usually due to a growth of the exposure, also the second risk factor, Credit/Volatility in our examples, must be hedged. If at the same time the trader considers it cost-effective to rebalance also the first exposure F , then hedging goes back to standard Black and Scholes recipe that sets local variance to zero. If instead the hedge in the second risk factor is rebalanced without a simultaneous rebalancement of the first hedge (that the trader could consider not cost effective) the optimal recipe can be obtained following the same variance minimization approach seen above.

General case

We assume the amount invested in F is a generic Δ_t^F and we look for $g_C(\Delta_t^F)$ minimizing $\text{Var}[d\Pi_t - \Delta_t^F dF_t - g_C dC_t]$, which corresponds to

$$\text{Var}\left[\left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right)dF + \left(\frac{\partial\Pi}{\partial C} - g_C\right)dC\right]$$

The C -hedge minimizing this quantity is

$$g_C(\Delta_t^F) = \frac{\partial\Pi}{\partial C} - \frac{\text{Covar}(dC, dF)}{\text{Var}(dC)} \left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right) = \boxed{\frac{\partial\Pi}{\partial C} + \rho \left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right) \frac{\sigma_F}{\sigma_C}}.$$

Since we could symmetrically start from hedging C and hedge F later, we have a symmetric definition for $g_F(\Delta_t^C)$, and we notice $g_F = g_F(0)$.

Again we obtain the same result if we perform a global, rather than local, variance minimization. We look for an adapted process g_C such that the self-financing strategy which at time t holds Δ_t^F of asset F and $g_C(t)$ of asset C (to be determined) has a discounted value $\bar{H}_t$ minimizing $\text{Var}[\bar{\Pi}_T - \bar{H}_T]$. By self-financing

$$\bar{H}_t = \bar{H}_0 + \int_0^t \Delta_s^F d\tilde{F}_s + \int_0^t g_C d\tilde{C}_s,$$

hence we look for $b := \left(\frac{\partial\Pi}{\partial C} - g_C\right)$ minimizing

$$\begin{aligned} \text{Var}\left[\int_0^T \left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right) d\tilde{F}_t + \int_0^T \left(\frac{\partial\Pi}{\partial C} - g_C\right) d\tilde{C}_t\right] = \\ \int_0^T \mathbb{E}\left[D(0, t)^2 \left(\left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right)^2 \sigma_F^2 + b^2 \sigma_C^2 + 2\rho\sigma_C\sigma_F b \left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right)\right)\right] dt. \end{aligned}$$

As in Section 1.2.1, pointwise minimization is possible: computations lead to

$$b = -\rho \left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right) \frac{\sigma_F}{\sigma_C}, \quad \text{i.e.} \quad g_C = \frac{\partial\Pi}{\partial C} + \rho \left(\frac{\partial\Pi}{\partial F} - \Delta_t^F\right) \frac{\sigma_F}{\sigma_C}.$$

This is again independent of the choice of the horizon T .

Special case

The result in the last equation is general. In the specific case that Δ_t^F equals g_F computed above (an approximation for the case when rebalancement of the C -hedge is not simultaneous to F -rebalancement, yet it happens a “short time” after it) we obtain a very simple solution, given by

$$\left(\frac{\partial \Pi}{\partial C} - g_C\right) = -\frac{\text{Covar}(dC, dF)}{\text{Var}(dC)} \left(-\frac{\sigma_C}{\sigma_F} \rho \frac{\partial \Pi}{\partial C}\right) \rightarrow \rho^2 \frac{\partial \Pi}{\partial C},$$

so our finding is

$$g_C(g_F) = (1 - \rho^2) \frac{\partial \Pi}{\partial C}.$$

For a larger number of correlated assets, the approach can be generalized to any correlation matrix (see Appendix A.1), although the special case of two assets makes it easier to grasp the underlying intuitions. Notice the algorithm is telling us that when $\rho = \pm 1$ there is only one hedge, which makes sense since the second asset’s movement is perfectly predicted by the first asset’s movement. Moreover, when $\rho = 0$ we have just classical hedging, which is what we expect.

Apart from these special cases, notice that even if we use both assets, the final hedge is not perfect (the only perfect hedge is classical delta hedging), and results in a residual variance (we skip computations)

$$\rho^2(1 - \rho^2) \left(\frac{\partial \Pi}{\partial C}\right)^2 \sigma_C^2 dt.$$

If at the beginning we had considered “easier” to hedge C rather than F , we would have obtained a residual variance

$$\rho^2(1 - \rho^2) \left(\frac{\partial \Pi}{\partial F}\right)^2 \sigma_F^2 dt.$$

In cases when the two assets are equivalent in terms of “easiness” of hedging, which may mean that they have equivalent liquidity conditions, and yet hedging is affected by transaction costs, a trader may desire in any case to design the partial hedging strategy seen so far. In this case the comparison of these two variances will decide which asset must be treated as “first” one.

If instead in a given moment we were given the possibility to rebalance both components of the hedging portfolio, the optimal choice would be, needless to say, the Black and Scholes hedge that sends to zero the local variance of the total portfolio, irrespectively of any correlations between the two assets.

These considerations are a first glimpse to the problem of formalizing, in order to make it more efficient, the decision process implemented by the trader when hedging in illiquid markets. The problem is analysed in the next section.

1.3 Realistic optimal hedging

In this section we design a hedging strategy that implements the above optimal hedging results while taking into account the constraints we have in the real world, and using only inputs available in practice. We start from an analysis that digs into the implications of utility maximization under transaction costs but we reach a formulation of the strategy that does not require utility parameters or detailed definitions of transaction costs.

1.3.1 Preliminary analysis of utility and transaction costs

The formalization of optimal hedging under transaction costs that had most impact on the literature is given in Hodges and Neuberger (1989), that approaches the problem by maximizing the expected value of the utility of the trader. Risk averse utility functions $U(w)$ must have $U'(w) > 0$ and $U''(w) < 0$, where w is the random amount of money whose utility needs to be assessed. Hodges and Neuberger (1989) make a classical choice, using a negative exponential utility

$$U(w) = -e^{-\lambda w}.$$

For $w \sim \mathcal{N}(M, V)$, where V indicates variance, knowledge of the characteristic function implies

$$\mathbb{E}[U(w)] = -e^{-\lambda M + \frac{1}{2}\lambda^2 V} \quad (1.4)$$

that clearly rewards positive expectation and penalizes variance. They get a stochastic optimal control problem characterized by a Bellman-Hamilton-Jacobi equation.

The approach of utility maximization appears correct, but, in this case as much as in many other financial applications, it adds complication without adding anything to usability. In fact utility maximization can be useful to formalize the problem, but any solution based on explicit representation of utility is not practical, since no one knows which utility function should be used, and which one is the correct value of its parameters. This means that any practical recipe must deduce implicitly the “utility function” of the trader by some actual decisions he makes in his activity.

In their practical activity, traders monitor the volatility of their P&L by monitoring their exposure, namely the difference between a) the sensitivity $\partial\Pi/\partial X$ of their derivative portfolio to the different risk factors, and b) the quantity Δ_t^x invested as a hedge in each one of the risk factors. In case of one single asset, the representation is simple. Traders monitor

$$\left. \frac{\partial\Pi}{\partial X} \right|_t - \Delta_t^x$$

and in this way, with or without the support of a model, but always using their practical knowledge of the volatility conditions of X , they monitor implicitly the local standard deviation of the P&L as

$$\left| \left. \frac{\partial\Pi}{\partial X} \right|_t - \Delta_t^x \right| \sigma_x \sqrt{\delta} = \sqrt{V_t} \quad (1.5)$$

where δ is a short interval of time. At some point $t = s$, they decide that they no longer accept the level of volatility reached by the global portfolio, and rebalance the hedge making

$$\Delta_{t=s}^x = \left. \frac{\partial\Pi}{\partial X} \right|_s$$

The moment s cannot depend only on the volatility of the portfolio, since we know that the level of volatility accepted as a consequence of mishedging is different from asset to asset, with a higher volatility tolerance for less liquid assets X .

As we pointed out in the introduction, this form of decision making we observe in practice is a consequence of the presence of fixed costs in real market trading. Unfortunately, most previous literature assumes linear/proportional transaction costs, in the sense that purchase or sale of quantity Δ of an asset with price X incurs transaction costs given by

$$c|\Delta|X,$$

see the review paper Zakamouline (2005). Under such transaction costs, continuous hedging is infeasible, but frequent rehedging is acceptable as long as one buys or sells only small amounts of underlying. This contradicts reality, and does not correspond to what happens in markets. Traders tend to minimize the re-hedging frequency, and, for all but the most liquid assets, they seem to consider inefficient to rehedg when rehedging implies buying or selling just a small quantity of assets. This is due to transaction costs having a fundamental fixed component. This fixed component of hedging costs is a synthesis for many real facts: the implicit cost of committing to get a good deal for the hedging instrument in an illiquid market, the existence of a more or less explicit “minimum amount” which is standard to buy or sell, and explicit fixed fees imposed by brokers. This fixed component of hedging costs is our starting point to design a strategy that minimizes hedging costs and yet allows some immunization against the movements of *all* risk factors even if only some are actually rebalanced at a given time. Considering both a linear and a fixed part could lead to a recipe not sufficiently simple and intuitive to be implemented in a dynamic activity like hedging, without giving any practically significant advantage.

The trader knows that rehedging has a fixed transaction cost c^x . Utility function allows to transform volatility of P&L into a monetary cost, in particular we can use the so called marginal rate of substitution to model precisely the trade-off between transaction costs and volatility. Namely, for a generic utility function, one can always use the implicit function theorem to compute the quantity

$$\frac{dM}{dV}$$

which is telling us which is the increase dM in the expectation of the random payout w that we need to receive if we want to keep utility unchanged in case of an increase of volatility dV . Moreover, if utility has the form (1.4), then dM/dV is constant and equal to $\lambda/2$: hence

$$\Delta M = \frac{\lambda}{2} \Delta V$$

holds also for any finite increment of volatility ΔV .

We call c^x the fixed cost of transacting in the underlying X . The trader will rebalance its sensitivity when

$$\frac{\lambda}{2} V_t$$

first touches c^x from below: in fact when the trader rebalances he reduces local volatility from V_t to zero, while he spends c^x . Thus, in consideration of (1.5), his hedging strategy acts at any time t in which $(X_t, \partial \Pi_t / \partial X)$ exits from the following “no-transaction region”:

$$\left(\left. \frac{\partial \Pi}{\partial X} \right|_t - \Delta_t^x \right)^2 \sigma_x^2 < \frac{2}{\lambda \delta} c^x := \Theta_x. \quad (1.6)$$

The strategy is completely determined by a single threshold value Θ_x , in which the effects due to risk-aversion (λ), utility horizon (δ) and transaction cost (c^x) are jointly taken into consideration without explicitly referring to the (unmeasurable) variables used in its derivation.

In practice, traders know when the cost of transacting overcomes the benefit from volatility reduction, and also know when the situation reverts. This trader behaviour is more directly described by an overall level Θ_x than by the implicit parameters λ , c and δ .

1.3.2 The possible hedging choices

Now suppose the trader's portfolio is exposed to the two risk factors F and C , and at time t he holds in its portfolio quantities Δ_t^F and Δ_t^C of the two assets. The cost of acting in F is c^F , while for C cost is c^C . Thanks to the results of the previous section, the trader knows that at any time t there are only four possible actions:

- 0 No rebalancing, since no re-hedging choice is cost-effective taking into account the trade-off between illiquidity and variance. There is obviously no variance reduction.
- 1 Rebalancing both Δ_t^F and Δ_t^C , with cost $c^F + c^C$. In this case, optimal rebalancement leads locally to B&S's perfect hedge. Hence the volatility reduction is given by the volatility of the total portfolio before rebalancement:

$$\Delta_1 V = V_{t-} = \delta \left\{ \sigma_F^2 \left(\Delta_{t-}^F - \frac{\partial \Pi}{\partial F} \right)^2 + \sigma_C^2 \left(\Delta_{t-}^C - \frac{\partial \Pi}{\partial C} \right)^2 + 2\rho\sigma_F\sigma_C \left(\Delta_{t-}^F - \frac{\partial \Pi}{\partial F} \right) \left(\Delta_{t-}^C - \frac{\partial \Pi}{\partial C} \right) \right\}. \quad (1.7)$$

- 2 Rebalancing only Δ_t^F , with cost c^F . Earlier in this chapter we showed that the optimal hedge when we act on Δ_t^F with Δ_t^C fixed, taking correlation into account, is

$$g_F(\Delta_t^C) = \frac{\partial \Pi}{\partial F} + \rho \frac{\sigma_C}{\sigma_F} \left(\frac{\partial \Pi}{\partial C} - \Delta_t^C \right); \quad (1.8)$$

we substitute this choice into the formula for the local variance and get

$$V_{t+} = (1 - \rho^2) \sigma_C^2 \left(\frac{\partial \Pi}{\partial C} - \Delta_t^C \right)^2 \delta,$$

leading to a variance reduction

$$\Delta_2 V = V_{t-} - V_{t+} = \delta \left\{ \sigma_F^2 \left(\Delta_{t-}^F - \frac{\partial \Pi}{\partial F} \right)^2 + \boldsymbol{\rho^2} \sigma_C^2 \left(\Delta_{t-}^C - \frac{\partial \Pi}{\partial C} \right)^2 + 2\rho\sigma_F\sigma_C \left(\Delta_{t-}^F - \frac{\partial \Pi}{\partial F} \right) \left(\Delta_{t-}^C - \frac{\partial \Pi}{\partial C} \right) \right\}. \quad (1.9)$$

The only difference from (1.7) is the term ρ^2 in boldface.

- 3 Rebalancing only Δ_t^C . Everything goes as in point 2, exchanging the role of F and C , so that

$$g_C(\Delta_t^F) = \frac{\partial \Pi}{\partial C} + \rho \frac{\sigma_F}{\sigma_C} \left(\frac{\partial \Pi}{\partial F} - \Delta_t^F \right), \quad (1.10)$$

and the variance reduction is

$$\Delta_3 V = V_{t-} - V_{t+} = \delta \left\{ \boldsymbol{\rho^2} \sigma_F^2 \left(\Delta_{t-}^F - \frac{\partial \Pi}{\partial F} \right)^2 + \sigma_C^2 \left(\Delta_{t-}^C - \frac{\partial \Pi}{\partial C} \right)^2 + 2\rho\sigma_F\sigma_C \left(\Delta_{t-}^F - \frac{\partial \Pi}{\partial F} \right) \left(\Delta_{t-}^C - \frac{\partial \Pi}{\partial C} \right) \right\}. \quad (1.11)$$

1.3.3 A strategy based solely on P&L volatility

Here we show how a choice can be made at any time. We introduce for convenience of exposition the auxiliary variables

$$f := \sigma_F \left(\Delta^F - \frac{\partial \Pi}{\partial F} \right), \quad c := \sigma_C \left(\Delta^C - \frac{\partial \Pi}{\partial C} \right). \quad (1.12)$$

Calling c_i the cost of each of the three reheding “moves” among which the trader can choose, the new utility after a reheding of the i -th type is

$$U_{t+} = -\exp \left\{ \frac{\lambda^2}{2} (V - \Delta_i V) - \lambda (M - c_i) \right\} = U_{t-} \cdot \exp \left\{ \lambda \left(c_i - \frac{\lambda}{2} \Delta_i V \right) \right\}$$

and therefore its attractiveness is higher if the multiplier of the negative quantity U_{t-} is lower. So, the better move is the one for which

$$G_i = \frac{\lambda}{2} \Delta_i V - c_i \quad (1.13)$$

is higher. No action is taken if $G_i < 0$ for all $i = 1, 2, 3$.

In consideration of (1.7), (1.9) and (1.11), we have that

$$\begin{aligned} G_1 &= \frac{\lambda}{2} \delta (f^2 + c^2 + 2\rho fc) - (c^F + c^C) = \frac{\lambda \delta}{2} [(f^2 + c^2 + 2\rho fc) - (\Theta_F + \Theta_C)], \\ G_2 &= \frac{\lambda}{2} \delta (f^2 + \rho^2 c^2 + 2\rho fc) - c^F = \frac{\lambda \delta}{2} [(f + \rho c)^2 - \Theta_F], \\ G_3 &= \frac{\lambda}{2} \delta (\rho^2 f^2 + c^2 + 2\rho fc) - c^C = \frac{\lambda \delta}{2} [(\rho f + c)^2 - \Theta_C]. \end{aligned} \quad (1.14)$$

The key observation is that since in these equations $(\lambda \delta)/2$ is a constant positive multiplier, the trader can assess the relative and absolute convenience of each of the three actions using only the quantities Θ_F , Θ_C defining his behaviour in the single-asset case.

The possible ambiguities are eliminated by the fact that under the most standard utility function the marginal rate of substitution is independent of the level of volatility. This is exemplified in the next section through an analysis of the geometry of hedging in the real world.

1.3.4 A graphical analysis of the geometry of hedging

After the initial setup of the hedge, in view of (1.14), the proposed strategy acts always to keep true the following system of inequalities:

$$\begin{cases} f^2 + c^2 + 2\rho fc < \Theta_F + \Theta_C, \\ (f + \rho c)^2 < \Theta_F, \\ (\rho f + c)^2 < \Theta_C, \end{cases} \quad (1.15)$$

which therefore describes the no-transaction region R . This is the intersection of:

- An ellipse $R_{F,C} = \{f^2 + c^2 + 2\rho fc < \Theta_F + \Theta_C\}$, whose axes are rotated by $\frac{\pi}{4}$ with respect to the f and c axes; the major axis is in the second and fourth quadrant if and only if $\rho > 0$, and the eccentricity grows with $|\rho|$.

When the boundary $\partial R_{F,C}$ of this ellipse is touched, a complete reheding is performed.

- The strip R_F between the two lines $\{f + \rho c = \pm\sqrt{\Theta_F}\}$. When either boundary of this strip is touched, a F -rehedging is performed with Δ^C fixed.
- The strip R_C between the two lines $\{\rho f + c = \pm\sqrt{\Theta_C}\}$. When either boundary of this strip is touched, a C -rehedging is performed with Δ^F fixed.

As for the rehedging actions, they have the following geometric interpretation:

1. Rebalancing both Δ^F and Δ^C moves (f, c) to the origin.
2. Rebalancing only Δ^F moves (f, c) horizontally to the line $\{f + \rho c = 0\}$ bisecting the strip R_F .
3. Rebalancing only Δ^C moves (f, c) vertically to the line $\{\rho f + c = 0\}$ bisecting the strip R_C .

This geometry allows a number of observations.

1. In the degenerate case $\rho = 1$, we have:

$$R_{F,C} = \{|f+c| < \sqrt{\Theta_F + \Theta_C}\}, R_F = \{|f+c| < \sqrt{\Theta_F}\}, R_C = \{|f+c| < \sqrt{\Theta_C}\}.$$

The complete B&S rehedging is never performed, since one can always achieve the complete variance reduction acting only on F or only on C , thanks to perfect correlation. The trader will always use F if $\Theta_F < \Theta_C$, and will always use C otherwise; i.e., he will always buy and sell only the cheapest hedging instrument, as common sense would suggest. (See Figure 1.2a.)

The case $\rho = -1$ is completely analogous (see Figure 1.2b).

2. When $|\rho| < 1$, the strategy is almost surely unambiguous, since almost every point of the boundary of R violates only one inequality in (1.15) and therefore triggers only one possible rebalancement. In fact, the intersections between any couple of sets among $\partial R_{F,C}$, ∂R_F and ∂R_C are always in finite number.
3. When $\rho = 0$, one *never* changes simultaneously Δ^C and Δ^F , because $R_F \cap R_C$ is a rectangle properly contained in the interior of $R_{F,C}$ (see Figure 1.2c). The strategy involves monitoring and rebalancing independently the positions in F and C as in the single-asset case.
4. Whichever the choice of the parameters, the strategy never collapses to using *always* the classical B&S hedge. In fact, the two points

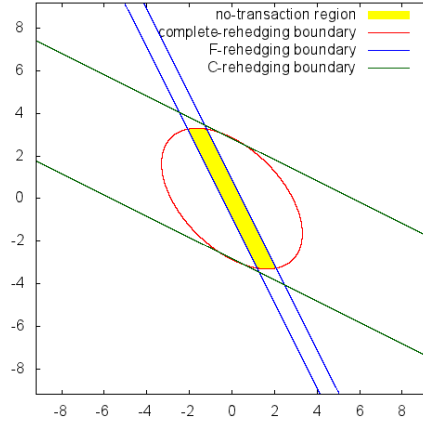
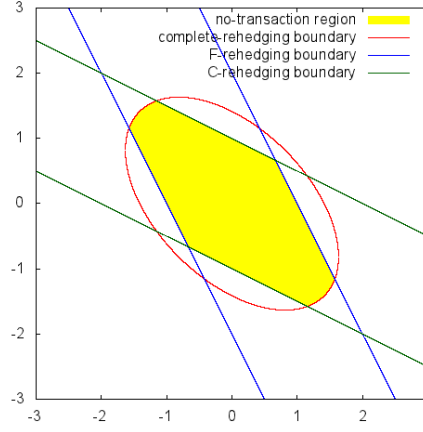
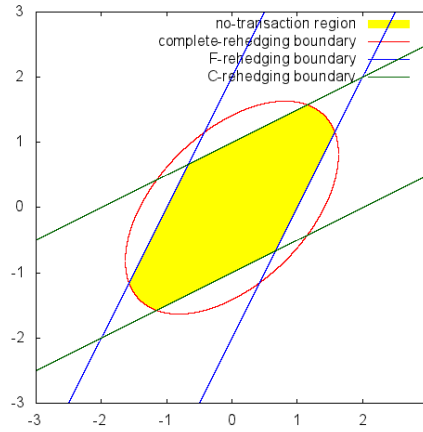
$$(f, c) = \left(\pm \min \left\{ \sqrt{\Theta_F}, \frac{\sqrt{\Theta_C}}{\rho} \right\}, 0 \right)$$

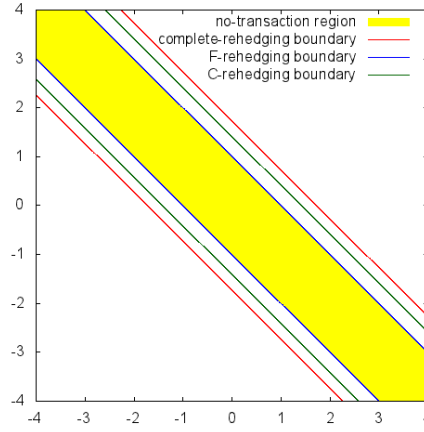
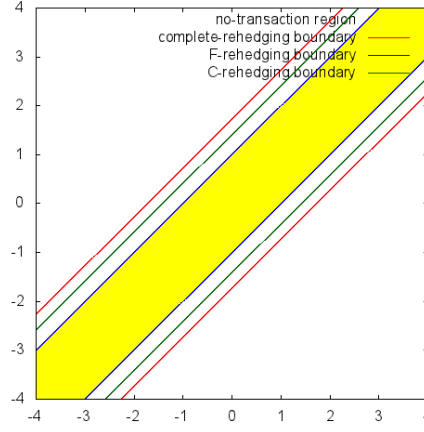
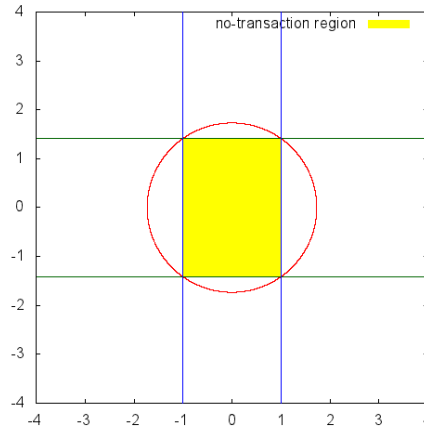
always belong to the boundary of R but to the interior of $R_{F,C}$: hence, in a neighborhood of those points in ∂R , a partial rehedging is performed.

5. When $\Theta_F \ll \Theta_C$, C -rehedging is never performed (see Figure 1.1a); when $\Theta_C \ll \Theta_F$, F -rehedging is never performed; while in general, the hedging strategy involves all the three possible “moves”, i.e. complete rehedging, F -rehedging and C -rehedging. (See Figure 1.1b.)

1.4 Practical case study

Now we design a practical test of the hedging strategy outlined above. The case study we have in mind is hedging of counterparty risk of an Interest Rate Swap, but we try to keep the setting as simple and general as possible, for the reader to appreciate and understand the results also in view of other applications. This implies first to keep the modelling framework simple.

(a) No-transaction region for $\rho = \frac{1}{2}, \Theta_F = \frac{1}{5}, \Theta_C = 8$.(b) No-transaction region for $\rho = \frac{1}{2}, \Theta_F = \Theta_C = 1$.(c) No-transaction region for $\rho = -\frac{1}{2}, \Theta_F = \Theta_C = 1$.Figure 1.1: No-transaction region in the (f, c) plane, for different choices of the parameters.

(a) No-transaction region for $\rho = 1$, $\Theta_F = 1$, $\Theta_C = 2$.(b) No-transaction region for $\rho = -1$, $\Theta_F = 1$, $\Theta_C = 2$.(c) No-transaction region for $\rho = 0$, $\Theta_F = 1$, $\Theta_C = 2$.Figure 1.2: No-transaction region in the (f, c) plane, for different choices of the parameters.

1.4.1 Model choice

We consider only two risk factors, both modelled as single-factor Hull&White diffusions (Hull and White, 1990), with a strong 80% correlation linking their stochastic drivers:

$$\begin{aligned} dr(t) &= k_r(\theta_r(t) - r(t))dt + \sigma_r dW_r(t), \\ dh(t) &= k_h(\theta_h(t) - h(t))dt + \sigma_h dW_h(t), \\ d\langle W_r, W_h \rangle_t &= \rho dt, \quad \rho = 80\%, \end{aligned}$$

where

- r is the money market short rate and will be used to price an interest rate swap maturing in 10 years;
- h plays the role of a default intensity or hazard rate in that the payoff Π will be expressed as

$$\text{CVA}(\text{Swap}) = \mathbb{E} \left[\int_0^T e^{-\int_0^t (h(s) + r(s)) ds} \text{Swap}(t, r_t)^+ dt \right]. \quad (1.16)$$

In the numerical results, we have kept the same volatility, mean reversion and initial point for the two processes, to avoid that differences in parameterizations could distort the comparison of the results of the different hedging strategies. Additionally, our parameterization is such that the $h(t)$ process keeps away from zero, as one expects for a default intensity. Yet, notice that keeping $h(t)$ Gaussian makes the setting applicable also to hedging of a swaption in some multicurve setting, where $h(t)$ could be interpreted as the negative spread that lowers the rate at which the payout is discounted, to reach the level consistent with a collateral agreement. Even more generally, we would expect similar results for any payoff in the form of an underlying multiplied by some discounting term $\exp\left(-\int_0^t x(s) ds\right)$.

1.4.2 The treatment of correlation

While sketchy in the model choices, we tried to be realistic in the treatment of correlation in two respects:

1. We supposed constant correlation *between two hidden risk factors*. This maps to a stochastic correlation between the two hedging instruments.
2. Although the paths of r and h were in fact produced by a simulation where the correlation parameter was chosen by ourselves, we pretended that we did not know it, and used part of the simulation to get a sample estimate of ρ . The estimate we obtain is 81.7%, slightly different from the actual correlation we assumed.

One may wonder why we are designing the example under the assumption that correlation will be historically estimated rather than being an implied correlation chosen for pricing. The reason for that has to do with the constraints that reality imposes to the theory underlying pricing and hedging. In reality, the models that are used for pricing in finance are very simple compared to the complexity of reality. The two one factor models above are a very clear example. In models with fewer factors than those existing in reality, modellers must make choices on which aspects of a multifaceted reality are most relevant. So, if two risks have to be correlated, one will look at the

type of correlation that has most influence in pricing, and this will be most often a long-term macro correlation. For example, when correlating a stock price with its stochastic volatility, say in a Heston (Heston, 1993) model, correlation will likely be negative, because this is what is needed to fit the market skew. But this correlation explaining the market skew is a long-term correlation, modelling the fact that when stock prices fall abruptly volatilities tend to jump in the opposite direction. Even if such movements have been seen only few times in decades, at the relevant market crises, yet this is the most important feature to get the right price. On the other hand, the short-term, standard comovements of underlying and volatility could be independent or even positively correlated, and these short-term comovements are what matters for day-by-day P&L volatility of a hedging strategy. In a more advanced model, the smile could be captured by jumps or by correlating some additional drivers of the long term means, leaving freedom to set diffusive correlations, and therefore hedging, to levels based on short term co-movements. But in a simple model one has to make a choice depending on the application one has in mind. For hedging purposes, the most convenient choice is historical correlation based on day-by-day comovements. A similar example may regard the credit risk of a heavily-indebted sovereign and the level of inflation of the currency in which debt is denominated. For large and long-term movements, it is logic to expect a negative correlation between the two, since high inflation reduces real public debt. But this does not imply at all that a slight increase in credit risk leads to slight reduction of inflation. Day-by-day movements can be influenced by many different facts.

1.4.3 Design of the strategies

The payoff we consider is the CVA of an in-the-money payer ten year maturity interest rate swap with nominal 1.000.000 Euros. Each day in the simulation, we took as available hedging instruments a par Interest Rate Swap and a Credit Default Swap, both maturing in 5 years. Note that (1.16) expresses the payoff as a function of r and s , hence we need to map the sensitivities $(\partial\Pi/\partial r, \partial\Pi/\partial h)$ given by this equation, to the couple $(\partial\Pi/\partial\text{Swap}, \partial\Pi/\partial\text{CDS})$ needed by our strategy. To this aim, it is sufficient to solve the linear system

$$\begin{cases} \frac{\partial\Pi}{\partial r} = \frac{\partial\Pi}{\partial\text{Swap}} \frac{\partial\text{Swap}}{\partial r} + \frac{\partial\Pi}{\partial\text{CDS}} \frac{\partial\text{CDS}}{\partial r}, \\ \frac{\partial\Pi}{\partial h} = \frac{\partial\Pi}{\partial\text{CDS}} \frac{\partial\text{CDS}}{\partial h}, \end{cases}$$

whose coefficients

$$\frac{\partial\text{Swap}}{\partial r}, \frac{\partial\text{CDS}}{\partial r}, \frac{\partial\text{CDS}}{\partial h}$$

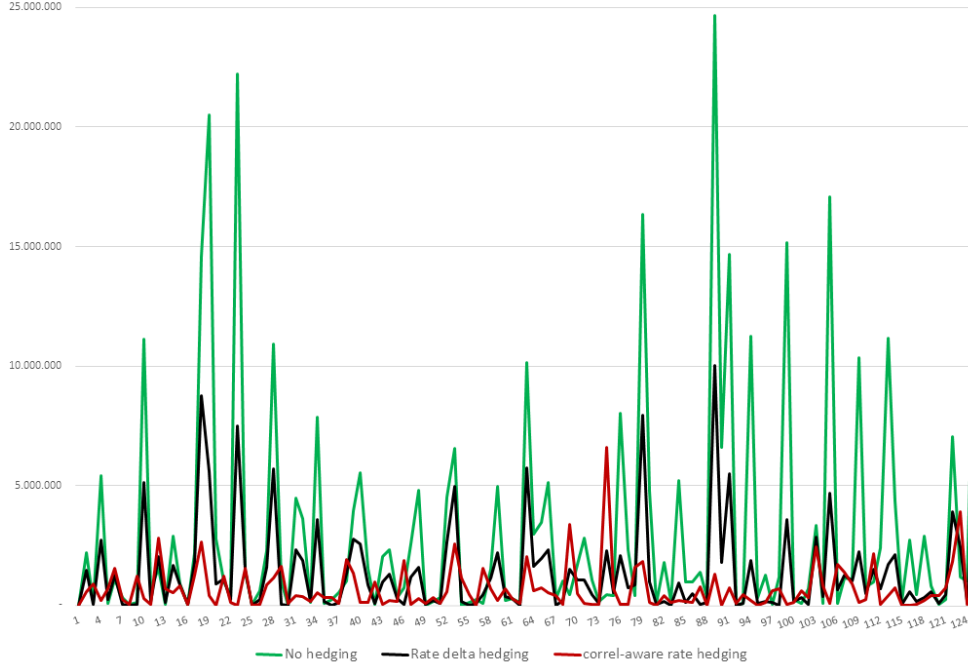
can be computed by the elementary pricing formulas of IRS and CDS.

We first run a simulation for 6 months assuming we do not hedge our exposure (**no hedging** in Table 1.1). The P&L volatility, computed as the quadratic average of wealth movements, turns out to be 1,798 Euros.

Now we assume that the trader has decided to hedge only one exposure, using the standard Black and Scholes approach for Delta-hedging, that is computing the chosen hedge as a simple, correlation unadjusted first derivative $g_X = \partial\Pi/\partial X$ to be recomputed every day based on the current value of the underlying (**single-hedge** in Table 1.1). If the exposure which is rebalanced is Interest Rates, this reduces P&L

Hedging instrument	Strategy	P&L volatility (Euros)
None	no hedging	1,797.81
IRS	single-hedge	1,187.57
IRS	single-hedge- ρ -aware	826.16
CDS	single-hedge	758.31
CDS	single-hedge- ρ -aware	522.29

Table 1.1: Comparison of portfolios reheded using a single instrument.

Figure 1.3: Daily reheding with single instrument: **no hedging**, **single-hedge**, and **single-hedge- ρ -aware**. Daily profit&loss (squared) as a function of time.

volatility to 1,188 Euros ($\sim 33\%$ improvement). If it is Credit, it reduces P&L volatility to 758 Euros ($\sim 58\%$ improvement).

Now we wonder what is the advantage in terms of P&L volatility if the trader uses the same approach of rebalancing daily but with the correlation-aware hedging formula given in (1.2) (**single-hedge- ρ -aware** in Table 1.1). Can it reduce volatility further? When only rates are hedged, using correlation-aware formulas reduces P&L volatility to 826 Euros; when only credit is hedged the correlation-aware formula leads to 522 Euros ($\sim$ further 31% improvement in both cases).

This shows that, even without the general hedging framework introduced in Section 1.3, just the correlation-aware formulas can give significant benefit to reducing P&L volatility. Results are summarized in Table 1.1. In Figure 1.3 instead we see graphically the evolution of the (portfolio + hedge) P&L in 3 different cases: no hedging, only hedging of one asset in the Black and Scholes way, and only hedging in the single-hedge- ρ -aware strategy.

Now we move to testing the framework introduced in Section 1.3, where also cost of re-hedging enters the picture and all hedging and no-hedging decisions of the trader are set in a rational context. For this second part of our tests, we need to distinguish the liquidity conditions of the two hedging products. We assume, realistically, that the swap is more liquid than the CDS, an assumption that in practice turns into a higher observable tolerance of the trader to P&L volatility of the credit exposure compared to the interest rate exposure, $\Theta_F \ll \Theta_C$. Looking at (1.6), in fact, we see that the cost c_i of rebalancing is the main driver of the difference in the rebalancing thresholds Θ_i .

We first consider standard daily Black and Scholes two-asset Delta hedging, where hedges, recomputed daily, are given by $g_F = \partial\Pi/\partial F$, $g_C = \partial\Pi/\partial C$ (**B&S hedging** in Table 1.2). As expected, this reduces P&L volatility by at least an order of magnitude compared to the above tests: only 49 Euros. But this comes with the cost of 250 rebalancing operations, half of them performed on the illiquid credit hedge with high rebalancing costs.

Then we run an experiment where we use the full hedging strategy designed in this chapter, where rebalancing decisions happen based on P&L volatility thresholds, and rebalancement of hedges takes correlation into account: the strategy described in 1.3.2 and 1.3.3 with hedges computed via (1.8) and (1.10) (**Θ & ρ -aware strategy** in Table 1.2). We expect fewer rebalancements, corresponding to a reduction of rehedging costs, possibly with some increase of P&L volatility. We have indeed an increase of P&L volatility: our strategy leads to a P&L volatility of 66 Euros, 17 more than in the Black and Scholes case. This minor increase in P&L volatility comes however with an impressive reduction in transaction costs: in fact, instead of 125 rebalances of swap hedging and 125 rebalances of CDS hedging, we only have 18 swap rebalances and, even more importantly, only 6 rebalances of the illiquid and thus costly CDS hedge. If, as we assumed above, CDS re-hedging is the dominant part of transaction costs, we have reduced transaction costs to 5% of the Black and Scholes figures.

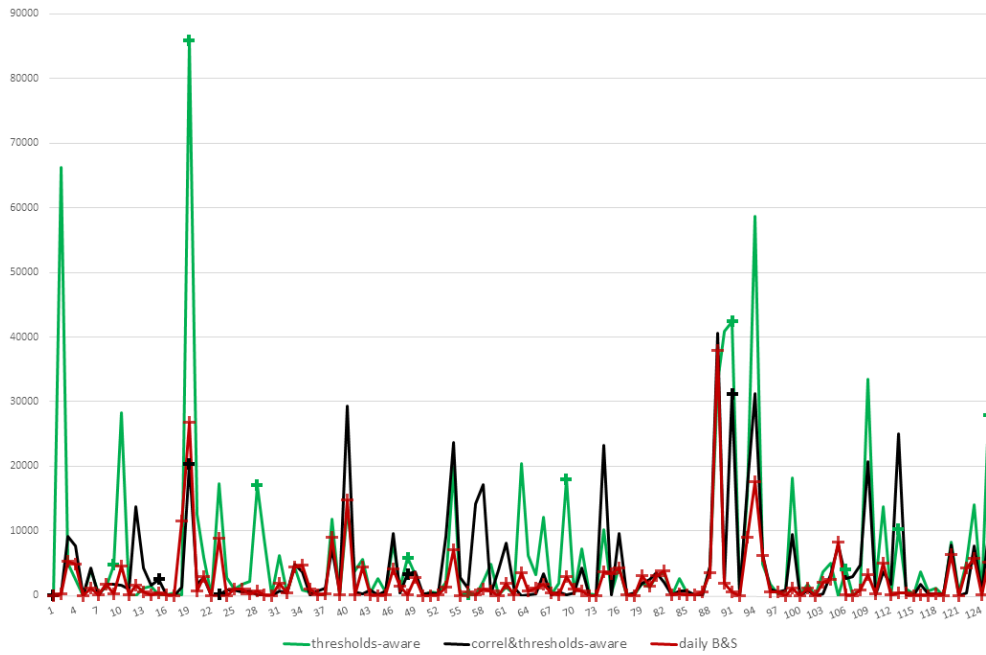
In order to understand this result better, we tried another test: we used the hedging strategy where rebalancing decisions happen based on P&L volatility thresholds, as described in 1.3.3, but we compute hedges using standard Black and Scholes Deltas $g_F = \partial\Pi/\partial F$, $g_C = \partial\Pi/\partial C$, neglecting correlation adjustments like those in (1.8) and (1.10) for both rates and credit (**Θ -aware-no- ρ strategy** in Table 1.2). This is enough to outperform dramatically pure Black and Scholes in term of rehedging costs, in fact we have only 16 swap rehedges vs 125 and 11 CDS rehedges vs 125, and the loss in terms of P&L volatility is minimal, considering that P&L volatility is now 82, just 33 Euros more than Black and Scholes and 16 Euros more than the Θ & ρ -aware strategy. Yet if, as we assumed above, CDS re-hedging is the dominant part of transaction costs, notice that the Θ & ρ -aware strategy still allows a further reduction of transaction costs by almost 50% compared to this Θ -aware strategy that neglects correlation adjustments.

The results of these tests are summarized in Table 1.2. In Figure 1.4 we see the behaviour of the portfolio P&L in the last three cases of Table 1.2. Since these are different strategies involving different rehedging frequencies, in this chart we have indicated the costly credit rehedges with little crosses.

This allows us to draw some conclusions. In single-hedging strategies, where one risk-factor is not hedged, due to technical restrictions or to the high cost of the hedge, the correlation-aware hedging formulas we provide are crucial to minimize P&L volatility. In multi-hedging strategies where all risk factors are hedged, the threshold-based strategy seems very useful even when correlation is not taken into account, since it gives a P&L volatility which is less than twice the B&S P&L volatility, but with an

Strategy	IRS rehedges	CDS rehedges	P&L vol (€)
no hedging	0	0	1,797.81
B&S hedging	125	125	49.26
Θ & ρ -aware	18	6	65.76
Θ -aware-no- ρ	16	11	82.07

Table 1.2: Comparison of hedging strategies using both underlying assets.

Figure 1.4: Reheding strategies with both assets: Θ -aware-no- ρ strategy, Θ & ρ -aware strategy and B&S hedging. Daily profit&loss (squared) as a function of time.

impressive reduction in terms of reheding costs (just 11 credit re-hedges in 6 months, as opposed to daily rehedges). The most impressive advantage, however, emerges in the complete strategy where correlation is taken into account. In the latter case, in fact, P&L volatility is just a fraction higher than B&S (66 vs 50), with just 6 credit re-hedges vs 125 re-hedges in the B&S strategy. We focus on credit re-hedges since they have been assumed to be the most expensive ones, but conclusions are not altered if we take into account also the savings in terms of interest rate re-hedges.

1.5 Conclusions

We have detailed the theory of exploiting correlation in hedging, showing that correlation among hedges enters explicitly the design of an optimal computation of Greeks in many practical contexts. We have also designed a well-defined trading strategy that, differently from most literature on hedging in discrete time and under transaction costs, does not require the knowledge of the parameters of the utility function of the trader. We have also provided a geometric interpretation.

We have finally shown experimentally that the strategy performs well in a CVA hedging numerical test where the risk factors follow standard highly-correlated diffusive models. Our approach reduces P&L volatility when the frequency of re-hedging is fixed, and when used also to determine re-hedging times it allows a massive reduction of re-hedging frequency, with relevant potential saving of hedging costs.

Useful extensions could be studied in the future. From the theoretical point of view, one may wonder whether there are different choices of utility function and transaction cost which lead to well-defined strategies, whether risk can be measured in subtler ways than bare short-term variance, or whether dependence can be defined otherwise than simple diffusion correlation. From an empirical point of view, the efficacy of the strategy may be tested in more problematic cases, such as multiple assets or different assumptions on dynamics, maybe up to a non-parametric approach where only historical data are used.

Chapter 2

Hedging under recalibration

Because of the need to recalibrate models frequently, the practice of hedging requires the computation of the sensitivities of prices directly with respect to the market quotes used in calibration. If calibration is not exact, this implies a loss of information in the passage from many quotes to few model parameters, and hence an indetermination in the concept of Greek. This chapter uses the language of differential geometry to formalize this ambiguity, and discusses the consequences on the construction of the hedging portfolio.

2.1 Introduction

In quantitative finance, the price of a financial product is almost always expressed as a function of the parameters defining a financial model, and the current value of its state variables. At least when the evolution of the state variables is supposed to be continuous, standard theory tells that all risks can be neutralized if one manages his portfolio in such a way that the first order derivative of its value with respect to each state variable is always equal to zero.

Nevertheless, it is common practice to compute and monitor also the sensitivity to model parameters. Indeed, model parameters are typically estimated from market quotes, i.e. numerically chosen so that the present available market information is reproduced at least approximately. This procedure is called calibration; in practice, it is repeated even daily, so that the “constants” which define the model are not constant at all in time.

To be clear: obviously, an imperfect calibration is sound evidence that the chosen model does not describe reality correctly. We add that even if we can calibrate perfectly, but at the price of making model parameters time-dependent, we are falsifying the constness of model parameters, an assumption which we have for sure used in computing the price. Hence, we have no clue on the coherence of our prices in terms of no-arbitrage.

Because of this unknown nature of the true dynamics which drive the movements of market quotes (an almost metaphysical concept), any attempt to state that an hedging strategy is optimal or even only good in some quantified sense is problematic: even if one has an optimization goal which is well-posed despite arbitrage, the optima will highly depend on the full probability distribution of the true dynamics, which we have no idea about (otherwise why did we use the wrong model for pricing in the first place?).

The limited theoretical literature addressing these issues concentrates on single-stock markets and postulates that the model misspecification is benign in an ad-hoc sense. For instance, to prove that misspecified Delta hedging is loss-free, El Karoui et al. (1998) need that the model is wrong only in that it overestimates volatility; recalibration is not considered. More recently, Herrmann and Muhle-Karbe (2017) have introduced in the picture (perfect) recalibration to a single option; their deep optimality result is only asymptotic in a factor penalizing a suitable “distance” of the true dynamics from Black-Scholes.

All the above suggests that:

- A satisfactorily general theoretical justification of hedging under recalibration is at very least not in sight, and we should even consider the possibility that there is no clean answer to this bigger problem;
- To date, the only general-purpose model-agnostic results are invariably based on sensitivities.

The second point is in line with market practice. In fact, in absence of better indications, practitioners usually consider as additional risk factors the market quotes themselves, whose movements are reflected into movements of calibrated parameters, and they try to neutralize directly the sensitivities of their portfolio to them. For the above reasons, this objective of neutralizing first order movements is not unreasonable at all, and *we take it for granted*. Instead, in this chapter we want to put some order and formalism in its application. We believe that here the whole point is the meaning that one wants to give to the word sensitivity in this context.

One first idea is to start computing first order “model Greeks” as derivatives of the pricing function with respect to model parameters. This is a well-studied problem, for which many general-purpose techniques have been proposed, for instance Malliavin calculus (Malliavin and Thalmaier, 2006) or algorithmic differentiation (Capriotti, 2011; Capriotti and Giles, 2012; Giles, 2007); see also the second part of the present thesis. Then the task becomes to translate this information into “market Greeks”, propagating derivatives through calibration.

The case in which the number of constraints imposed by observation of the market is equal to the number of parameters to be calibrated, in such a way that any configuration is exactly reproducible with a proper choice of them, is satisfactorily treated in Henrard (2011). The key remark is that, at least locally, the mapping of quotes into parameters is usually bijective and differentiable, hence model Greeks and market ones are essentially a reparametrization of each other.

However, the case in which there are less model parameters than market constraints is more delicate. One could in principle treat the full pricing algorithm inclusive of calibration like any other functional relationship, and compute its derivatives. This corresponds to the industry’s bump-and-revalue algorithm, in which the calibration+pricing routine is re-run after a small displacement of a single quote to approximate the sensitivity of the price to it as a difference quotient. It is also the point of view of Henrard (2013), which proposes an efficient (but a priori questionable) approximation of the Jacobian of the calibration to quickly convert model Greeks into market Greeks. The results of such methods will strongly depend on arbitrary choices such as that of calibration weights: we wonder whether we can say something more intrinsic on sensitivities to market quotes without relying on a specific calibration algorithm.

To this purpose, it is worth remembering that the computation of sensitivities is aimed at monitoring market risks and at hedging them via the construction of a hedging portfolio. In practice, the choice of a model which is parsimonious in

parameters often corresponds with an hedging strategy built using a limited number of instruments. Strictly speaking, such strategies immunize the position only from *some* movements of the market. For instance, if the financial product depends on more than one interest rate, sometimes the only Greek considered is the one corresponding to *parallel* movements of the rate curve, in which all rates move of the same quantity. The choice looks somewhat arbitrary, and we would deem more sensible to hedge the market movements *which are compatible with the model*: if the model is somewhat reasonable, they will be more similar to the most plausible evolutions.

Indeed, every model defines implicitly a set of market configurations which it is able to replicate. This circumstance can be interpreted as a set of implicit constraints among quotes, and in this form it is well represented in terms of differential geometry. The use of this language allows to explore and understand the interactions among the rigidity of the model, the resulting ambiguity in the concept of Greek (including under a new light a close relative to the existing methodologies), and eventually the implied flexibilities in the choice of the hedging portfolio, which could help in the everyday practice of imperfect hedging.

The rest of the chapter is structured as follows. Section 2.2 formalizes the problem of the interaction between hedging and recalibration, starting from a nonstandard view of traditional concepts of Greeks to introduce the fundamental equation (G); Section 2.3 discusses its indetermination. The equation is solved in Sections 2.4 and 2.5, first in a particular case of wide applicability, then in an abstract and fairly general context; the solution turns out to be of very easy implementation. Section 2.6 is devoted to formalization of the concept of hedge in our framework, and Section 2.7 compares our methodology to market practice. All the machinery is then taken into action in Sections 2.8 and 2.9, the former hinting at possible practical uses of the different solutions of (G), the latter presenting some numerical results. The last section is devoted to conclusions and possible developments.

2.2 Formalization of the problem

2.2.1 Model calibration

A continuous-time model of the marketplace is typically described by a set of stochastic processes

$$(S_1(t), \dots, S_d(t)) = \mathbf{S}(t), \quad (X_1(t), \dots, X_s(t)) = \mathbf{X}(t), \quad t \in [0, T], \quad d, s \geq 0,$$

where the S_i are tradable assets, while the $\mathbf{X}$ vector represents some model-dependent state.¹ For instance, each S_i may be a stock price, and X_i the corresponding instantaneous variance in a stochastic volatility model. The dynamics of $(\mathbf{S}, \mathbf{X})$ are parametrized by real numbers

$$(\alpha_1, \dots, \alpha_a) = \boldsymbol{\alpha};$$

in the stochastic volatility case, these may be the vol-of-vol, mean-reversion-of-vol, long-term-vol, and instantaneous correlation parameters.

¹The condition of tradability of S_i could be easily relaxed, requiring only the availability of a sufficiently rich set of functions of the observable indices S_i ; we avoid this generality to simplify the exposition.

Suppose that in addition to the S_i , some non-directly-modelled market instruments are available for trading, with quotes

$$q_{d+1}(t), \dots, q_n(t).$$

After the instrument with index $i > d$ is subscribed at time t_0 , its NPV will fluctuate because of the movements of the corresponding quote and of the underlying market $\mathbf{S}$, according to some known function

$$H_i(t, \mathbf{S}(t), q_i(t)), \quad i > d$$

of current time, visible state and quote.

We underline that the function H_i is a matter of market convention: for example, if the market instruments are call options and q_i are implied volatilities, then H_i is the Black pricing formula for the i -th option just by definition of implied volatility. This does not force us to choose a lognormal model: under our model, the price of the i -th instrument will be a different function

$$\phi_i(t, \boldsymbol{\alpha}, \mathbf{X}(t), \mathbf{S}(t))$$

of time, model parameters and state.

The information shipped by quotation $q_i(t)$ translates into a model identity

$$f_i(t, \boldsymbol{\alpha}, \mathbf{X}(t), \mathbf{S}(t), q_i(t)) = 0, \quad (2.1)$$

for instance putting

$$f_i(t, \boldsymbol{\alpha}, \mathbf{X}(t), \mathbf{S}(t), q_i(t)) = \phi_i(t, \boldsymbol{\alpha}, \mathbf{X}(t), \mathbf{S}(t)) - H_i(t, \mathbf{S}(t), q_i(t)).$$

These equation are used to *calibrate* analytically or numerically the model, i.e. to determine the value of $\boldsymbol{\alpha}$ and $\mathbf{X}(t)$.

To ease notation in some subsequent equations, we introduce symbols to denote all quantities which are given and all quantities which are calibrated:

$$\begin{aligned} \mathbf{y}(t) &= (\mathbf{S}(t), \mathbf{q}(t)) \in \mathbb{R}^n, \boldsymbol{\theta}(t) = (\boldsymbol{\alpha}, \mathbf{X}(t)) \in \Omega, \\ \mathbf{f}(t, \boldsymbol{\theta}(t), \mathbf{y}(t)) &= (f_i(t, \boldsymbol{\alpha}, \mathbf{X}(t), \mathbf{S}(t), q_i(t)))_{i=d+1, \dots, n} \in \mathbb{R}^k, \end{aligned}$$

where Ω is an open subset of $\mathbb{R}^m$.² In the sequel, we fix a time t to drop the t arguments, and denote by Π_θ , Π_y the canonical projections $(\boldsymbol{\theta}, \mathbf{y}) \mapsto \boldsymbol{\theta}$ and $(\boldsymbol{\theta}, \mathbf{y}) \mapsto \mathbf{y}$ respectively. We define analogously $\Pi_S(\mathbf{S}, \mathbf{q}) = \mathbf{S}$ and $\Pi_q(\mathbf{S}, \mathbf{q}) = \mathbf{q}$.

The concept of calibration which is more widely used in practice is the following:

Definition 2.1 (Bestfit calibration). We will call **bestfit calibration** a function

$$\tilde{c} : \mathbf{y} \mapsto \boldsymbol{\theta}^{\min} \text{ such that } \boldsymbol{\theta}^{\min} \in \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=d+1}^n f_i^2(\boldsymbol{\theta}, \mathbf{y}).$$

We note that $\tilde{c}$ can be defined for each $\mathbf{y} \in \mathbb{R}^n$, but in general $\mathbf{f}(\tilde{c}(\mathbf{y}), \mathbf{y}) \neq \mathbf{0}$.³ In this case, $\mathbf{y}$ is rigorously speaking incompatible with the chosen model. For this reason, we provide a second definition:

²The openness of the domain simplifies the subsequent mathematics considerably; we deem useful and interesting also the case in which Ω is a regular closed set, but we leave out this case for future study.

³This would be true even for more complex calibration procedures, based for instance on stochastic filtering for the estimation of the hidden factors $\mathbf{X}$. We stick to least-square calibration for simplicity of exposition.

Definition 2.2 (Exact calibration). We will call **exact calibration** a function

$$c : \mathbf{y} \mapsto \boldsymbol{\theta}^{\text{zero}} \text{ such that } \mathbf{f}(\boldsymbol{\theta}^{\text{zero}}, \mathbf{y}) = \mathbf{0}.$$

It can be defined only for $\mathbf{y} \in Y$, where

$$Y = \Pi_{\mathbf{y}} M, \quad M = \{(\boldsymbol{\theta}, \mathbf{y}) : \mathbf{f}(\boldsymbol{\theta}, \mathbf{y}) = \mathbf{0}\}.$$

Although restrictive in asking for a perfectly-calibratable input $\mathbf{y} \in Y$, in a dynamical setting this kind of calibration is still more generous than a strict application of our model would allow. Indeed, if the model is a perfect description of reality, once one has inferred $\boldsymbol{\theta}(0) = (\boldsymbol{\alpha}, \mathbf{X}(0))$ from exact calibration at time 0, the following calibration should be sufficient:

Definition 2.3 (In-the-model calibration). For each $\boldsymbol{\alpha}$, we will call **in-the-model calibration given $\boldsymbol{\alpha}$** a function

$$c^{\boldsymbol{\alpha}} : \mathbf{y} \mapsto \mathbf{X}^{\text{zero}} \text{ such that } \mathbf{f}(\boldsymbol{\alpha}, \mathbf{X}^{\text{zero}}, \mathbf{y}) = \mathbf{0}.$$

It can be defined only for $\mathbf{y} \in Y^{\boldsymbol{\alpha}}$, where

$$Y^{\boldsymbol{\alpha}} = \{\mathbf{y} \in Y : \mathbf{f}(\boldsymbol{\alpha}, \mathbf{X}, \mathbf{y}) = \mathbf{0} \text{ for some } \mathbf{X}\}.$$

Remark 2.4. We note that c is the restriction of $\tilde{c}$ to Y , and $c^{\boldsymbol{\alpha}}$ is a projection of the restriction of c to $Y^{\boldsymbol{\alpha}}$. Moreover, one can see $c^{\boldsymbol{\alpha}}$ as an instance of c for a different model: precisely, the model with no free parameters ($a = 0$) obtained from the original one fixing a priori $\boldsymbol{\alpha} = \boldsymbol{\alpha}(0)$ and thus dropping the explicit dependence on $\boldsymbol{\alpha}$ in all functions.

Example. Consider an hypothetical one-dimensional model with a single hidden factor $X = \theta \in \mathbb{R}$ (i.e. $s = m = 1$, $a = d = 0$), whose value we wish to calibrate to 2 market instruments (i.e. $k = n = 2$), whose prices y_1, y_2 are quoted. Suppose that according to the model, the first price should be $\cos \theta$ and the second one should be $\sin \theta$.

Then in our formalism:

$$\begin{cases} f_1(\theta, \mathbf{y}) = \cos \theta - y_1, \\ f_2(\theta, \mathbf{y}) = \sin \theta - y_2. \end{cases}$$

The sets M and Y are drawn in Figure 2.1.

2.2.2 Hedging

Now suppose that a trader has in his book a financial product, and at a given time t denote by

$$p(\mathbf{S}, \boldsymbol{\theta})$$

its pricing function according to the model, which will be assumed to be sufficiently smooth. Let us say broadly that the trader wants to hedge market risk, which gives non-deterministic behaviour to the evolution of the value, adding to his book a (dynamically rebalanced) portfolio $\mathcal{S}$ in the underlying market $\mathbf{S}$ and another portfolio $\mathcal{H}$ in the hedging instruments $\mathbf{H}$. What does this mean exactly? As discussed in the introduction, there are at least two radically different ways of interpreting the task:

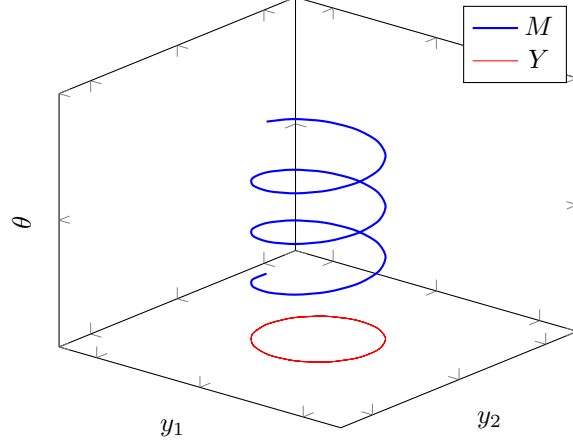


Figure 2.1: In the Example, $M = \{(\theta, y_1, y_2) : y_1 = \cos \theta, y_2 = \sin \theta\}$ is a spiral, and the set of the $\mathbf{y}$ which are compatible with the model is its projection $Y = \{(y_1, y_2) : y_1^2 + y_2^2 = 1\}$, i.e. the unit circumference.

1. *The textbook trader.* If a trader has a firm belief in the correctness of the model, then he knows that at least if $\mathbf{X}$ and $\mathbf{S}$ are continuous semi-martingales, he can achieve perfect hedging by designing $\mathcal{S}$ and $\mathcal{H}$ in such a way that

$$\frac{\partial}{\partial \mathbf{S}} (p + \mathcal{S} + \mathcal{H}) = 0, \quad \frac{\partial}{\partial \mathbf{X}} (p + \mathcal{S} + \mathcal{H}) = 0. \quad (2.2)$$

Note, as we will discuss in detail in the sequel, that the hedging portfolio might be non-unique.

2. *The skeptical trader.* Aware of the (incoherent) practice of recalibration, real-life traders usually consider the pricing function from market quotes

$$\tilde{P}(\mathbf{S}, \mathbf{q}) = p(\mathbf{S}, \tilde{c}(\mathbf{S}, \mathbf{q})) = (p \circ (\Pi_{\mathbf{S}}, \tilde{c}))(\mathbf{y})$$

as a whole, and try to put to zero its derivatives:

$$\frac{\partial}{\partial \mathbf{y}} (\tilde{P} + \mathcal{S} + \mathcal{H}) = 0. \quad (2.3)$$

In their current display, (2.2) and (2.3) are difficult to compare, but we claim that they can be rewritten in a common form, which will suggest a halfway alternative.

Indeed, if one sticks to (2.2), everything he needs about the financial product is the sensitivity to model state $\partial p / \partial (\mathbf{S}, \mathbf{X})$; but traders tend to dislike the information in this form, because its mapping to the composition of the hedging portfolio is nontrivial. To understand why, suppose that $\partial p / \partial X_i$ is non-null only for one fixed i : the trader can conclude nothing on “where” the sensitivity is, since in principle all hedging instruments depend on all X_j -s, and he does not know a-priori which and how many of them he will need to bring to zero the single $\partial p / \partial X_i$.

Instead, the trader would rather have a set $(g_i)_{i=1, \dots, n}$ of “sensitivities” to market input $\mathbf{y}$, so that it will be trivial for him to match $(g_i)_{i>d} = \mathbf{g}_H$ with the portfolio $\mathcal{H}$ buying an amount of the i -th hedging instrument given by the familiar hedge ratio

$$r_i = -g_i \left[\frac{\partial H_i}{\partial q_i} \right]^{-1}, \quad i > d, \quad (2.4)$$

and then to match the “Delta” $(g_i)_{i \leq d} = \mathbf{g}_S$ of the resulting portfolio $P + \mathcal{H}$, buying an amount of S_i given by

$$r_i = -g_i - \sum_{j=d+1}^n r_j \frac{\partial H_j}{\partial S_i}, \quad i \leq d. \quad (2.5)$$

For the skeptical trader, these g_i have an immediate interpretation as properly-said sensitivities:

Proposition 2.5 (Skeptical trader’s sensitivities). *If the hedging portfolio composition is defined by (2.4)-(2.5), then the constraint (2.3) holds if and only if $\mathbf{g} = \text{grad } \tilde{P}$, i.e.*

$$d_{\mathbf{y}} \tilde{P}[\delta \mathbf{y}] = \sum_{i=1}^n g_i \delta y_i \quad \forall \delta \mathbf{y} \in \mathbb{R}^n, \quad (2.6)$$

where $d_{\mathbf{y}} \tilde{P}$ denotes the differential of function $\tilde{P}$ in point $\mathbf{y}$.

Proof. Note that (2.3) is in fact a system of n equations, the i -th of which zeroes $\partial/\partial y_i$ of the full portfolio. Substituting (2.4)-(2.5) in the i -th equation yields easily $g_i = \partial \tilde{P}/\partial q_i$ for $i > d$ and then $g_i = \partial \tilde{P}/\partial S_i$ for $i \leq d$, i.e. the thesis $\mathbf{g} = \text{grad } \tilde{P}$. $\square$

We claim that also in the case of (2.2), the g_i can be interpreted as sensitivities of the pricing functional as a function of market inputs, provided that in defining this functional one uses the stricter concept of in-the-model calibration instead of the broader bestfit one:

$$P^\alpha = p \circ (\Pi_S, c^\alpha).$$

To give a precise meaning to the claim, we should take into consideration two things:

1. We want derivatives to be defined for a function P^α which is defined only on the set of admissible inputs Y^α ;
2. The original equation (2.2) involves derivatives in all possible directions $(S_i)_{i=1,\dots,d}$ and $(X_i)_{i=1,\dots,s}$, which must therefore be spanned by infinitesimal displacements of the input $\mathbf{y}$ of P^α .

Therefore, the natural hypothesis is:

Hypothesis 2.6. Y^α is a differentiable submanifold of $\mathbb{R}^n$, and (Π_S, c^α) is differentiable at $\mathbf{y}$ with surjective differential onto $\mathbb{R}^d \times \mathbb{R}^s$.

Sufficient conditions will arise from the subsequent treatment (see Theorem 2.16 and Remark 2.17); for the moment we content ourselves with assuming that the hypothesis holds. Then we get the following analogous of Proposition 2.5:

Proposition 2.7 (Textbook trader’s sensitivities). *If the hedging portfolio composition is defined by (2.4)-(2.5), and Hypothesis 2.6 is satisfied, then the constraint (2.2) holds if and only if*

$$d_{\mathbf{y}} P^\alpha[\delta \mathbf{y}] = \sum_{i=1}^n g_i \delta y_i \quad \forall \delta \mathbf{y} \in T_{\mathbf{y}} Y^\alpha, \quad (2.7)$$

where $T_{\mathbf{y}} Y^\alpha$ denotes the tangent space to Y^α in point $\mathbf{y}$.

Proof. Equation (2.2) can be written in differential form as

$$dp + \sum_{i=1}^d r_i dS_i + \sum_{i=d+1}^n r_i d\phi_i = 0, \quad (2.8)$$

where all differentials are with respect to $(\mathbf{S}, \mathbf{X})$, and we recall that ϕ_i denotes the price of the i -th hedging instrument as a function of $(\mathbf{S}, \mathbf{X})$. Now (2.8) is unaltered if we compose it to the right with $d(\Pi_S, c^\alpha)$, since this is onto by hypothesis: recalling that $\phi_i \circ (\Pi_S, c^\alpha) = H_i$, we get

$$dP^\alpha + \sum_{i=1}^d r_i dS_i + \sum_{i=d+1}^n r_i dH_i = 0$$

where all differentials are intended as defined on $T_{\mathbf{y}}Y^\alpha$. Substituting in this equation (2.4)-(2.5) and

$$dH_i = \sum_{j=1}^d \frac{\partial H_i}{\partial S_j} dS_j + \frac{\partial H_i}{\partial q_i} dq_i$$

gives after some simplifications

$$dP^\alpha - \sum_{i=1}^d g_i dS_i - \sum_{i=d+1}^n g_i dq_i = 0$$

as functions defined on $T_{\mathbf{y}}Y^\alpha$, which applied to the generic $\delta\mathbf{y} \in T_{\mathbf{y}}Y^\alpha$ yields (2.7). $\square$

Comparing (2.6) to (2.7), we see that the core difference is the following:

1. The skeptical trader wants $\mathbf{g}$ to represent the first-order movement of the pricing functional from market data for *all* infinitesimal market changes $\delta\mathbf{y} \in \mathbb{R}^n$; to do this, he must rely on a quite arbitrary extension $\tilde{P}$ of the pricing functional to model configurations $\mathbf{y} \notin Y$ incoherent with the model, given by his calibration algorithm.
2. Instead, the textbook trader just cares about the movements $\delta\mathbf{y} \in T_{\mathbf{y}}Y^\alpha$ compatible with the model with *fixed* parameters α .⁴

It is now quite natural to note that if we want to consider more realistically that α will most probably change, the model itself can tell us something more than (2.7) about the g_i , intrinsically and without any recourse to an out-of-the-model calibration algorithm. This is achieved considering the pricing functional

$$P = p \circ (\Pi_S, c),$$

which has a larger domain than P^α in that it allows for changes in the model parameters, while being still completely specified by the model, contrarily to $\tilde{P}$.

Again, since the function P is defined only on Y , to give a meaning to derivatives of P the most natural hypothesis is the following one:

Hypothesis 2.8. Y is a differentiable submanifold of $\mathbb{R}^n$, and P is a differentiable function $Y \rightarrow \mathbb{R}$.

⁴Note that at least if the market is complete and the model is correct, these would be the only movements compatible with no-arbitrage.

In the next sections we will find sufficient conditions for this to be true; for the moment we content ourselves with assuming it, so that we can define the following compromise definition between the textbook and skeptical concepts of Greeks:

Definition 2.9 (Greeks equation). We will say that $\mathbf{g}$ satisfies the **Greeks equation** if the following holds:

$$d_{\mathbf{y}}P[\delta\mathbf{y}] = \sum_{i=1}^n g_i \delta y_i \quad \forall \delta\mathbf{y} \in T_{\mathbf{y}}Y. \quad (\text{G})$$

Morally, we are looking for quantities g_i such that if we move $\mathbf{y}$ by a small quantity $\Delta\mathbf{y}$, then the variation of P satisfies

$$\Delta P \approx \sum_{i=1}^n g_i \Delta y_i; \quad (2.9)$$

which is less stringent than it may seem, because not all $\Delta\mathbf{y}$ are admissible, being constrained by $\mathbf{y} + \Delta\mathbf{y} \in Y$. When we send to zero sequences of $\Delta\mathbf{y}$ which are admissible in this sense, the limit directions are the elements of $T_{\mathbf{y}}Y$, and we retrieve the infinitesimal formal version (G).

In the following sections we proceed studying the master equation (G); recall that by Remark 2.4 the results which will be obtained can be applied also to the theoretical-hedging equation (2.7).

2.3 Indetermination

The unstated implicit constraint $\mathbf{y} + \Delta\mathbf{y} \in Y$ in (2.9), which translates into the explicit constraint $\delta\mathbf{y} \in T_{\mathbf{y}}Y$ in (G), implies non-uniqueness of the solution $\mathbf{g}$, as the following example shows.

Example (continuation). In the context of the previous Example, let us suppose that according to the chosen model a financial product has value $p(\theta) = \sin^2 \theta$. Then $P(\mathbf{y}) = y_2^2$, hence a solution to our problem (2.9) is

$$(g_1, g_2) = \left(\frac{\partial[y_2^2]}{\partial y_1}, \frac{\partial[y_2^2]}{\partial y_2} \right) = (0, 2y_2).$$

However, the function P is defined only on Y , where the relation $y_1^2 + y_2^2 = 1$ holds: therefore $P(\mathbf{y}) = 1 - y_1^2$, so that another solution of our problem (2.9) is

$$(g_1, g_2) = \left(\frac{\partial[1 - y_1^2]}{\partial y_1}, \frac{\partial[1 - y_1^2]}{\partial y_2} \right) = (-2y_1, 0).$$

Moreover, any convex combination of these two solutions is itself a solution.

Remark 2.10 (Canonical representatives). The right hand side of (G) is the scalar product $\langle \mathbf{g}, \delta\mathbf{y} \rangle$: the only constraint on $\mathbf{g}$ is its scalar product with the elements of $T_{\mathbf{y}}Y$. As a consequence, it is immediate to verify that

$$\mathbf{g} \text{ solves (G)} \iff \forall \Delta\mathbf{g} \perp T_{\mathbf{y}}Y, \mathbf{g} + \Delta\mathbf{g} \text{ solves (G)}. \quad (2.10)$$

This characterizes the freedom in the choice of $\mathbf{g}$. Some plausible choices are:

1. A solution belonging to $T_{\mathbf{y}}Y$. Equation (2.10) implies immediately that such a solution exists unique, and that it is characterized by having the smallest Euclidean norm $\|\mathbf{g}\|_2$. This solution is the gradient of P in the sense of Riemannian geometry, therefore we will call it **Riemannian** Greek. More flexibly, one may minimize some nonnegative quadratic form of the g_i other than the bare sum of squares, obtaining **generalized Riemannian** Greeks: see Remark 2.11 for a geometric interpretation.
2. A choice of $\mathbf{g}$ with at most $\dim(Y)$ non-zero components. The existence follows from (2.10) with linear algebra arguments. This solution corresponds to the construction of an hedging portfolio with the minimal number of instruments, therefore we will call it **economical** Greek. Again, one can devise more flexible variants, for instance allowing $g_i \neq 0$ for i in a desired set larger than the minimal size, and using the remaining degrees of freedom to perform some kind of minimization in the spirit of point 1.
3. The gradient of $\tilde{P}$. It is a solution because function $\tilde{P}$ extends P , hence $d_{\mathbf{y}}\tilde{P}$ extends $d_{\mathbf{y}}P$, and therefore Proposition 2.5 implies immediately (G) with $\mathbf{g} = \text{grad } \tilde{P}$. This solution involves the bestfit calibration $\tilde{P}$, hence we will call it **bestfit** Greek.

The three choices are drawn in Figure 2.2 for our Example.

Remark 2.11 (Coordinate changes and rescalings). Instead of minimizing $\|\mathbf{g}\|_2^2$, one may minimize some nonnegative quadratic form of the g_i . The result lacks the nice geometric characterization involving the tangent space, but this can be interpreted as having chosen the “wrong” coordinate system.

Indeed, if one chooses linearly transformed coordinates $\mathbf{y}' = \mathbf{M}^{-1}\mathbf{y}$, then the minimized quantity will be

$$\|\mathbf{g}'\|^2 = \|\mathbf{M}^\top \mathbf{g}\|^2 = \mathbf{g}^\top (\mathbf{M} \mathbf{M}^\top) \mathbf{g},$$

which as $\mathbf{M}$ varies represents the generic nonnegative quadratic form of the g_i . We conclude that we can get a rather rich family of generalized Riemannian Greeks from standard (Euclidean) Riemannian Greeks with appropriate linear changes of coordinates.

The coordinate change is particularly reasonable when matrix $\mathbf{M}$ is diagonal, so that it just rescales market inputs y_i without mixing them. In fact, if the quantities with respect to which we are computing the derivatives are not homogeneous among themselves (for instance if they include some stock prices and some implied volatilities), to make the norm of $\mathbf{g}$ have dimensionally sense we should anyway rewrite the objective function $\mathbf{f}$ in terms of rescaled adimensional variables. Here we are only adding that one should choose the unit of measure for y_i in such a way to reflect its relevance in the objective function. See Remark 2.23 for a transparent application.

To solve concretely (G), we will need to write

$$dP = d_{(S, c(\mathbf{y}))} P \circ (\Pi_S, d_{\mathbf{y}} c).$$

The differential of the function $p : \mathbb{R}^d \times \mathbb{R}^m \rightarrow \mathbb{R}$ can be written with no difficulties as

$$dp = \sum_{j=1}^m \frac{\partial p}{\partial S_j} dS_j + \sum_{j=1}^m \frac{\partial p}{\partial \theta_j} d\theta_j,$$

where

$$\frac{\partial p}{\partial S_j} = \bar{S}_j, \quad \frac{\partial p}{\partial \theta_j} = \bar{\theta}_j$$

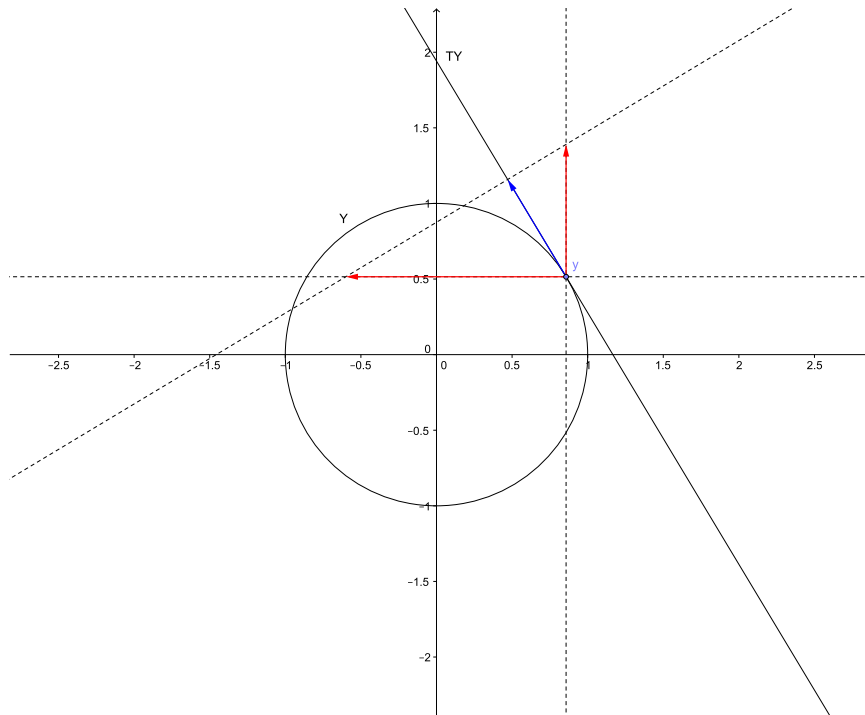


Figure 2.2: Graphical representation of some significant choices for a solution $\mathbf{g}$ of (G), in the Example we studied till now. You can see manifold Y (the circumference) and space $T_y Y$ (the tangent line). All possible solutions lie on the diagonal dashed line, which is orthogonal to $T_y Y$. The blue vector denotes the Riemannian Greek. We will see in Section 2.7 that for the particular form of $\mathbf{f}$ in our example, the blue solution coincides also with the bestfit Greek $\hat{P}$. The solutions in red are the two possible economical Greeks.

are the model Greeks, which we suppose to be well defined and at our disposal; we collect them in row vectors $\bar{\mathbf{S}}$ and $\bar{\boldsymbol{\theta}}$. It remains to understand whether c is in fact differentiable and to find its differential, in order to eventually solve (G). In the next section we will perform this analysis in a simple case of wide applicability; the most general case is the subject of Section 2.5.

2.4 A simple case

Suppose that $\mathbf{S}$ is empty. Then the quoted price of the i -th calibration instrument depends solely on $\boldsymbol{\theta}$, and a sensible form of the objective function would be

$$f_i(\boldsymbol{\theta}, \mathbf{y}) = \phi_i(\boldsymbol{\theta}) - y_i,$$

where $\phi_i(\boldsymbol{\theta})$ represents the price according to the model, while y_i is the price as quoted by the market.

We have $Y = \phi(\Omega)$.⁵ Therefore, to hope that Y is a manifold, the natural hypothesis is that the Jacobian $D\phi$ is full-rank.⁶ Then ϕ , restricted to a neighbourhood of $\boldsymbol{\theta}$, is a diffeomorphism with its image. Let us shrink Ω so that it is contained in such a neighbourhood: this implies no loss of generality, since to hope that P is differentiable, it is very reasonable to ask at least for continuity of c . In this way Y is a diffeomorphic image of an open subset of $\mathbb{R}^m$, hence a differentiable manifold of dimension m . Now we can eventually solve (G):

Theorem 2.12 (Solution of (G), simple case). *Suppose that $\mathbf{S}$ is empty, that each f_i is of the form $f_i(\boldsymbol{\theta}, \mathbf{y}) = \phi_i(\boldsymbol{\theta}) - y_i$ where $\phi_i \in C^1(\Omega)$, and that $D\phi$ is full-rank. Then, possibly shrinking Ω , Y is a differentiable manifold of dimension m , Hypothesis 2.8 is satisfied, and problem (G) is equivalent to the linear system*

$$\mathbf{g}^\top \cdot D\phi = \bar{\boldsymbol{\theta}}. \quad (2.11)$$

Proof. We already remarked that shrinking Ω we can suppose that Y is a differentiable manifold of dimension m . If we think ϕ as a map of Ω onto Y , we note immediately that $c = \phi^{-1}$, from which P is a composition of differentiable functions: Hypothesis 2.8 is satisfied.

We have to compute

$$\bar{\mathbf{y}} = d_{c(\mathbf{y})}p \circ d_{\mathbf{y}}c = \bar{\boldsymbol{\theta}} \circ d_{\mathbf{y}}(\phi^{-1}) = \bar{\boldsymbol{\theta}} \circ (d_{c(\mathbf{y})}\phi)^{-1},$$

i.e. the only $\bar{\mathbf{y}} \in T_{\mathbf{y}}^*Y$ such that $\bar{\mathbf{y}} \circ d_{c(\mathbf{y})}\phi = \bar{\boldsymbol{\theta}}$.⁷ If we recall that $\bar{\boldsymbol{\theta}} = \sum_i \bar{\theta}_i d\theta_i$ and that we wanted to write $\bar{\mathbf{y}}$ in the form $\sum_{i=1}^n g_i dy_i$, this equation becomes immediately (2.11). $\square$

Remark 2.13 (Algorithmic consequences). Once computed matrix $D\phi$ (say with algorithmic differentiation techniques), it remains to solve system (2.11), which as we expected is undetermined: the algorithm to compute $(g_1, \dots, g_n)$ will depend on the desired representative.

⁵Recall that in our notation, Ω is the open set where $\boldsymbol{\theta}$ can be chosen.

⁶This hypothesis is reasonable. Think about the limit case in which $\phi(\boldsymbol{\theta})$ does not depend on a certain θ_i : we could not hope to calibrate the model in an univocal way.

⁷We denote by $T_{\mathbf{y}}^*Y$ the cotangent space to Y in $\mathbf{y}$, i.e. the dual of the tangent space $T_{\mathbf{y}}Y$.

1. The Riemannian Greek corresponds to the minimal-norm solution. Since we have reduced our problem to a undetermined linear system, this solution is obtained with classical methods, such as singular value decomposition. Alternatively, one can exploit the subsequent Theorem 2.22. We remark that in a particular model setting, and mainly on the ground of empirical considerations, Piterbarg (2003) already suggests to use as Greeks the minimal-norm solution of an undetermined system which is essentially (2.11).
2. The economical Greek is obtained solving the system given by an $m \times m$ invertible submatrix of $D\phi$, and putting to zero the other components of $\mathbf{g}$. This corresponds to pretending, only for the purpose of this differentiation step, that we have calibrated only to $m < n$ instruments.
3. For the bestfit Greek $\tilde{P}$, we postpone the analysis of the algorithms to the beginning of Section 2.7.

2.5 The general case

We recall that $\boldsymbol{\theta} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{f}(\boldsymbol{\theta}, \mathbf{y}) \in \mathbb{R}^k$. Again we need differential geometry tools, hence we make the following weak hypothesis.

Hypothesis 2.14. $\mathbf{f}$ is C^1 and $D\mathbf{f}$ has full rank k , so that $M = \{\mathbf{f} = \mathbf{0}\}$ is a differentiable submanifold of $\mathbb{R}^{m+n}$ with dimension $m+n-k = m+d$. In particular, $m+n \geq k$.

All our argument needs obviously that $\mathbf{y} \mapsto \boldsymbol{\theta} = c(\mathbf{y})$ is a differentiable well-defined map. This means that we need a differentiable inverse of the restriction to set M of the canonical projection $\Pi_{\mathbf{y}} : (\boldsymbol{\theta}, \mathbf{y}) \mapsto \mathbf{y}$, which restriction we will denote by $\Pi_{\mathbf{y}|M}$. Therefore, it is natural to add the hypothesis that the differential of this map is full-rank, i.e.:

Hypothesis 2.15. $\Pi_{\mathbf{y}|T_{(c(\mathbf{y}), \mathbf{y})}M}$ has full rank $m+d$. In particular, $n \geq m+d$.

This looks non-trivial to check, hence we will give sufficient conditions (Theorem 2.18); for the moment, let us suppose that the hypothesis holds. Then, as in Section 2.4, restricting without loss of generality to a neighbourhood of $(c(\mathbf{y}), \mathbf{y})$ in M , we can suppose that $\Pi_{\mathbf{y}|M}$ is a diffeomorphism with its image Y , so that the latter turns out to be a differentiable manifold of dimension $m+d$.

The solution of (G) in this context is given by the following Theorem.

Theorem 2.16 (Solution of (G), general case). *Suppose that Hypotheses 2.14 and 2.15 hold. Then, restricting M to a neighbourhood of $(c(\mathbf{y}), \mathbf{y})$, Y is a differentiable manifold of dimension $m+d$, Hypothesis 2.8 is satisfied, and $\mathbf{g} = (\mathbf{g}_S, \mathbf{g}_q)$ solves (G) if and only if*

$$\begin{pmatrix} -\bar{\boldsymbol{\theta}}^\top \\ \mathbf{g}_S - \bar{\mathbf{S}}^\top \\ \mathbf{g}_q \end{pmatrix} \in (T_{(c(\mathbf{y}), \mathbf{y})}M)^\perp = \text{span}(\text{grad } f_i)_{i \leq k}. \quad (2.12)$$

Proof. We already saw that if we shrink M , then Y is a differentiable manifold of dimension $m+d$. We note that

$$c = \Pi_{\boldsymbol{\theta}|M} \circ \left(\Pi_{\mathbf{y}|M} \right)^{-1} \quad (2.13)$$

is a differentiable map, hence by composition $P = p \circ (\Pi_S, c)$ is itself differentiable, so that Hypothesis 2.8 is satisfied.

Our goal is to compute

$$\bar{\mathbf{y}} = d_{(\mathbf{S}, c(\mathbf{y}))} p \circ (\Pi_S, d_{\mathbf{y}} c) = \bar{\mathbf{S}} \circ \Pi_S + \bar{\boldsymbol{\theta}} \circ \Pi_{\theta|_{T(c(\mathbf{y}), \mathbf{y})M}} \circ \left(\Pi_{\mathbf{y}|_{T(c(\mathbf{y}), \mathbf{y})M}} \right)^{-1},$$

i.e. the only $\bar{\mathbf{y}} \in T_{\mathbf{y}}^* Y$ such that

$$\bar{\boldsymbol{\theta}} \circ \Pi_{\theta|_{T(c(\mathbf{y}), \mathbf{y})M}} = (\bar{\mathbf{y}} - \bar{\mathbf{S}} \circ \Pi_S) \circ \Pi_{\mathbf{y}|_{T(c(\mathbf{y}), \mathbf{y})M}}.$$

If we write $\bar{\mathbf{y}}$ in the desired form $\sum_i g_i dy_i$, $\bar{\mathbf{S}}$ as $\sum_i \bar{S}_i dS_i$ and $\bar{\boldsymbol{\theta}}$ as $\sum_i \bar{\theta}_i d\theta_i$, this equation becomes

$$\sum_{i=1}^m \bar{\theta}_i v_i = \sum_{j=1}^n g_j v_{m+j} - \sum_{i=1}^d \bar{S}_i v_{m+i} \quad \forall (v_1, \dots, v_{m+n}) \in T_{(c(\mathbf{y}), \mathbf{y})M},$$

i.e. exactly $\begin{pmatrix} -\bar{\boldsymbol{\theta}}^\top \\ \mathbf{g}_S - \bar{\mathbf{S}}^\top \\ \mathbf{g}_q \end{pmatrix} \in (T_{(c(\mathbf{y}), \mathbf{y})M})^\perp.$ □

Remark 2.17 (Surjectivity of differentials). Adding to the theorem the reasonable hypothesis that $(\Pi_S, \Pi_\theta)|_{T(c(\mathbf{y}), \mathbf{y})M}$ is surjective, i.e. that first-order-admissible market moves span all directions in $(\mathbf{S}, \boldsymbol{\theta})$ space, we can deduce from (2.13) that $d(\Pi_S, c)$ is surjective. Combined with Remark 2.4, this gives a way to verify Hypothesis 2.6.

Algorithmically, after computing $D\mathbf{f}$, one must solve the $(m+n)$ -dimensional linear system

$$\begin{pmatrix} -\bar{\boldsymbol{\theta}}^\top \\ \mathbf{g}_S - \bar{\mathbf{S}}^\top \\ \mathbf{g}_q \end{pmatrix} = \sum_{i=1}^k \lambda_i (\text{grad } f_i) \quad (2.14)$$

for the $n+k > n+m$ unknowns $g_1, \dots, g_n, \lambda_1, \dots, \lambda_k$. Also in this case, the solutions are in infinite number, and we can impose further conditions. For instance:

1. The minimum-norm solution is the one such that

$$\mathbf{g} \in T_{\mathbf{y}} Y = \Pi_{\mathbf{y}} [T_{(c(\mathbf{y}), \mathbf{y})M}] = \Pi_{\mathbf{y}} \left[(\text{span}(\text{grad } f_i)_i)^\perp \right].$$

In other terms, there must exist a $\boldsymbol{\theta}' \in \mathbb{R}^m$ such that

$$D\mathbf{f} \cdot \begin{pmatrix} \boldsymbol{\theta}' \\ \mathbf{g} \end{pmatrix} = \mathbf{0}. \quad (2.15)$$

Adding this k -dimensional equation and the m unknowns $\theta'_1, \dots, \theta'_m$ to system (2.14), we obtain a linear system with $m+n+k$ equations and $m+n+k$ unknowns.

We proceed as promised with a practical condition to check Hypothesis 2.15.

Theorem 2.18 (Sufficient conditions for Hypothesis 2.15). *Suppose that $n \geq k \geq m$, $\mathbf{f}$ is a C^1 function, $D\mathbf{f}$ has full rank k , and $D_{\boldsymbol{\theta}}\mathbf{f}$ has full rank m . Then Hypotheses 2.14 and 2.15 are satisfied, so that Theorem 2.16 can be applied.*

Proof. We must verify only Hypothesis 2.15. The submatrix $D_{\boldsymbol{\theta}}\mathbf{f}$ has rank $m \leq k$ where k is the rank of the full matrix $D\mathbf{f}$, hence by elementary linear algebra we can find a subset $\mathbf{y}^1$ of variables $\mathbf{y}$ such that $D_{(\boldsymbol{\theta}, \mathbf{y}^1)}\mathbf{f}$ is a $k \times k$ invertible matrix. Let

us write $\mathbf{y} = (\mathbf{y}^1, \mathbf{y}^2)$ with $\mathbf{y}^1 \in \mathbb{R}^{k-m}$ and $\mathbf{y}^2 \in \mathbb{R}^{n+m-k}$. By the implicit function theorem, there exists a differentiable function ψ defined on an open subset of $\mathbb{R}^{m+n-k}$ such that in a neighbourhood of $(c(\mathbf{y}), \mathbf{y})$, manifold M coincides with

$$\{(\boldsymbol{\theta}, \mathbf{y}^1, \mathbf{y}^2) : (\boldsymbol{\theta}, \mathbf{y}^1) = \psi(\mathbf{y}^2)\}.$$

Let us introduce the following notation to denote the components of ψ :

$$\psi_{\boldsymbol{\theta}} := \Pi_{\boldsymbol{\theta}} \circ \psi, \quad \psi_{\mathbf{y}^1} := \Pi_{\mathbf{y}^1} \circ \psi.$$

Then

$$\Psi : \mathbf{y}^2 \mapsto (\psi_{\boldsymbol{\theta}}(\mathbf{y}^2), \psi_{\mathbf{y}^1}(\mathbf{y}^2), \mathbf{y}^2)$$

is a diffeomorphism with its image, hence $\Pi_y|_{T_{(c(\mathbf{y}), \mathbf{y})}M}$ is full-rank if and only if $\Pi_y \circ d\Psi = d(\Pi_y \circ \Psi)$ is full-rank. To sum up, Hypothesis 2.15 is equivalent to the fact that the Jacobian matrix of

$$\Pi_y \circ \Psi : \mathbf{y}^2 \mapsto (\psi_{\mathbf{y}^1}(\mathbf{y}^2), \mathbf{y}^2)$$

should have full rank $m+n-k$. But a submatrix of this Jacobian is the identity $(m+n-k) \times (m+n-k)$: the conclusion follows. $\square$

Remark 2.19 (Coherence with the simple case). If $\mathcal{S}$ is empty and $f_i(\boldsymbol{\theta}, \mathbf{y}) = \phi_i(\boldsymbol{\theta}) - y_i$, the equations of this section should reduce to the ones in Section 2.4. This is not difficult to check: with $(e_i)_i$ denoting the canonical basis of $\mathbb{R}^n$, equation (2.14) becomes

$$\begin{pmatrix} -\bar{\boldsymbol{\theta}}^\top \\ \mathbf{g} \end{pmatrix} = \sum_{i=1}^k \lambda_i \begin{pmatrix} \text{grad } \phi_i \\ -e_i \end{pmatrix}, \quad \text{i.e.} \quad \begin{cases} -\bar{\boldsymbol{\theta}}^\top = \sum_i \lambda_i \text{grad } \phi_i, \\ g_j = -\lambda_j \quad \forall j. \end{cases}$$

If we substitute the second line of the system into the first one, we get $\bar{\boldsymbol{\theta}}^\top = \sum_i g_i \text{grad } \phi_i$, which is precisely (2.11).

If in addition we desire the minimum-norm solution, the additional condition (2.15) translates into

$$\exists \boldsymbol{\theta}' \in \mathbb{R}^m : \quad \sum_{j=1}^m \theta'_j \frac{\partial \phi_i}{\partial \theta_j} - g_i = 0 \quad \forall i \leq n,$$

which means exactly $\mathbf{g} \in \text{span} \left(\frac{\partial \phi}{\partial \theta_j} \right)_j$.

2.6 Linearity of Greeks and hedged portfolios

As we said, any solution of (G) has the right to be called a set of Greeks. For instance, one may desire to choose $\mathbf{g}$ in such a way to minimize some quantity other than the Euclidean norm $\|\mathbf{g}\|_2$ (maybe the total notional). However, the criteria suggested above (Remark 2.10) identify concepts of Greek with the desirable characteristic to be linear: the Greeks of a portfolio are the sums of the Greeks of the components. Such a condition is relatively simple to check for a given candidate algorithm, because it boils down to linearity of the map $(\mathcal{S}, \boldsymbol{\theta}) \rightarrow \mathbf{g}$ transforming model Greeks into market Greeks.

The demand for linearity deserves an in-depth consideration. Indeed, the typical goal of the sensitivities calculation is the construction of a hedging portfolio, such that if one adds it to the initial portfolio, the overall Greeks will be null.

In light of the analyses above, we will first of all need to clarify in which sense we want the Greeks to be “null”:

Definition 2.20 (Essential hedge). We will say that a portfolio is **essentially hedged** if $\mathbf{g} = \mathbf{0}$ is a possible solution of the Greeks equation (G).

Conceptually, a portfolio is essentially hedged if and only if at first order its value does not change for all variations of inputs *admissible according to the model*. Given the degrees of freedom in the choice of $\mathbf{g}$, having some $g_i \neq 0$ does not necessarily imply that the position is not hedged: on a purely theoretical ground, the only condition to be tested is $\mathbf{g} \perp T_{\mathbf{y}}Y$. However, one can note that a portfolio is essentially hedged if and only if $(\bar{\mathbf{S}}, \bar{\boldsymbol{\theta}}) = \mathbf{0}$, from which it is easy to conclude that for the three specific concepts of Greek examined in this work, the portfolio is essentially hedged if and only if $\mathbf{g} = \mathbf{0}$.

Therefore, it becomes strategic to have an algorithm converting a generic $\mathbf{g}$ into one of the canonical representatives of Remark 2.10). Indeed, it will not be uncommon that the starting point $\mathbf{g}$ may have been obtained as the sum of Greeks computed in different ways, for different purposes. For instance:

- The sensitivities of “primitive” hedging instruments $\mathbf{H}$ might have been computed analytically (say with a Black formula);
- The sensitivities of some deals, for which the trader has a precise idea of which instruments are more appropriate for hedging, might have been computed with the “economical” approach;
- The sensitivities of the remaining portfolio might have been computed in Riemannian sense (to have a minimal representation), or as the gradient of the bestfit calibration (the closest to the traditional concept of Greeks).

As an illustration, we show an algorithm to get the Riemannian Greeks from a generic $\mathbf{g}$, since (possibly after a change of metric as in Remark 2.11) it may be a good choice in the task of monitoring how much the portfolio is imbalanced, as we will briefly discuss in Section 2.8.

Proposition 2.21 (Conversion into Riemannian Greeks). *Let $\mathbf{g}$ be any solution of (G). Then the Riemannian solution $\mathbf{g}_R$ can be obtained solving simultaneously*

$$\begin{pmatrix} \mathbf{0} \\ \mathbf{g}_R - \mathbf{g} \end{pmatrix} \in \text{span}(\text{grad } f_i)_{i \leq k} \quad \text{and} \quad D\mathbf{f} \cdot \begin{pmatrix} \boldsymbol{\theta}' \\ \mathbf{g}_R \end{pmatrix} = \mathbf{0}.$$

for the unknowns $\mathbf{g}_R, \boldsymbol{\theta}'$.

Proof. The second equation is the condition of tangency as expressed in (2.15). The first equation can be derived noting that by Theorem 2.16, the solutions of (G) are obtained solving the undetermined linear system (2.12); notoriously the generic solution of such a system can be expressed as the sum of a single solution $\mathbf{g}$ (no matter which one) and the generic solution of the homogeneous system, i.e. the system obtained putting $\bar{\mathbf{S}} = \mathbf{0}$ and $\bar{\boldsymbol{\theta}} = \mathbf{0}$ in (2.12). $\square$

2.7 Comparison to the derivatives of bestfit calibration

We repeat briefly the argument in Henrard (2013) in order to adapt it to our context. The key remark is that $\boldsymbol{\theta}^{\min} = \tilde{c}(\mathbf{y})$ satisfies $\mathbf{F}(\boldsymbol{\theta}, \mathbf{y}) = \mathbf{0}$, where

$$\mathbf{F} = \text{grad}_{\boldsymbol{\theta}} \left[\frac{1}{2} \sum_{i=1}^k f_i^2 \right] = \sum_{i=1}^k f_i \text{grad}_{\boldsymbol{\theta}} f_i$$

is a function of $\mathbb{R}^m$ into itself. If $D_{\boldsymbol{\theta}}\mathbf{F}$ is invertible, then locally there is only one $\boldsymbol{\theta}$ verifying such relation, therefore this $\boldsymbol{\theta}$ must coincide with $\boldsymbol{\theta}^{\min}$. The conclusion is that $\tilde{c}$ is the implicit function defined by $\mathbf{F} = \mathbf{0}$, and therefore

$$D\tilde{c} = -(D_{\boldsymbol{\theta}}\mathbf{F})^{-1}(D_{\mathbf{y}}\mathbf{F}),$$

where by the definition of $\mathbf{F}$

$$\begin{cases} D_{\boldsymbol{\theta}}\mathbf{F} = \sum_i (\text{grad}_{\boldsymbol{\theta}} f_i)(\text{grad}_{\boldsymbol{\theta}} f_i)^{\top} + \sum_i f_i \text{Hess}_{\boldsymbol{\theta}} f_i, \\ D_{\mathbf{y}}\mathbf{F} = \sum_i (\text{grad}_{\boldsymbol{\theta}} f_i)(\text{grad}_{\mathbf{y}} f_i)^{\top} + \sum_i f_i (D_{\mathbf{y}} \text{grad}_{\boldsymbol{\theta}} f_i). \end{cases} \quad (2.16)$$

As a consequence, if one is able to compute the expressions above, and the model Greeks $\bar{S}_i, \bar{\theta}_j$ are known, one has all the ingredients to calculate the bestfit Greeks

$$(\text{grad } \tilde{P})^{\top} = (\bar{S}, \mathbf{0}) + \bar{\boldsymbol{\theta}} \cdot D\tilde{c} = (\bar{S}, \mathbf{0}) - \bar{\boldsymbol{\theta}} \cdot (D_{\boldsymbol{\theta}}\mathbf{F})^{-1}(D_{\mathbf{y}}\mathbf{F}). \quad (2.17)$$

We desire to understand whether there is some connection between these Greeks and the ones computed with our approach. Let us treat separately the cases $\mathbf{y} \in Y$ and $\mathbf{y} \notin Y$.

2.7.1 If $\mathbf{y} \in Y$

In this case, in both equations of (2.16) the second summation is zero, and what remains is

$$\begin{cases} D_{\boldsymbol{\theta}}\mathbf{F} = (D_{\boldsymbol{\theta}}\mathbf{f})^{\top}(D_{\boldsymbol{\theta}}\mathbf{f}), \\ D_{\mathbf{y}}\mathbf{F} = (D_{\boldsymbol{\theta}}\mathbf{f})^{\top}(D_{\mathbf{y}}\mathbf{f}). \end{cases}$$

Note that the condition of invertibility of $D_{\boldsymbol{\theta}}\mathbf{F}$ translates into $D_{\boldsymbol{\theta}}\mathbf{f}$ having full rank.

The relations above, in the simple case subject of Section 2.4, have an interesting consequence:

Theorem 2.22 (Sometimes Riemannian = bestfit). *Suppose that $f_i(\boldsymbol{\theta}, \mathbf{y}) = \phi_i(\boldsymbol{\theta}) - y_i$ under the hypotheses of Theorem 2.12. Then the minimum-norm solution of problem (G) is $\text{grad } \tilde{P}$.*

Proof. Equation (2.17) becomes

$$(\text{grad } \tilde{P})^{\top} = -\bar{\boldsymbol{\theta}} \cdot ((D\phi)^{\top}(D\phi))^{-1}(D\phi)^{\top}(-\mathbf{I}_n) = \bar{\boldsymbol{\theta}} \cdot ((D\phi)^{\top}(D\phi))^{-1}(D\phi)^{\top}. \quad (2.18)$$

The fact that $\text{grad } \tilde{P}$ is a solution of (G) is always true (see Section 2.2); anyway, Theorem 2.12 tells us that in the case we are now considering, equation (G) translates into $(\text{grad } \tilde{P})^{\top} D\phi = \bar{\boldsymbol{\theta}}$, which using (2.18) is a straightforward verification.

Finally, as we saw in Section 2.2, $\text{grad } \tilde{P}$ is the minimum-norm solution if and only if $\text{grad } \tilde{P} \in T_{\mathbf{y}}Y = \text{span}(\partial\phi/\partial x_i)_i$, which is true because in (2.18) $(\text{grad } \tilde{P})^{\top}$ is written as $\mathbf{a}^{\top} \cdot (D\phi)^{\top}$ for a vector $\mathbf{a}$. $\square$

What is the geometric interpretation of the above result? In the hypotheses of the theorem, the objective function of bestfit calibration $\sum_i (\phi_i(\boldsymbol{\theta}) - y_i)^2$ is squared distance of the theoretical point $\phi(\boldsymbol{\theta})$ from the market point $\mathbf{y}$, hence the calibrated point will be the projection (i.e. nearest point) of $\mathbf{y}$ on Y . In this setting, if $\mathbf{y}$ makes a small move orthogonal to Y , and one performs a new calibration, then the value of $\boldsymbol{\theta}$ does not change, implying no change in $\tilde{P}$: this means that $\langle \text{grad } \tilde{P}, \Delta \mathbf{y} \rangle$ must be null for every $\Delta \mathbf{y} \perp T_{\mathbf{y}}Y$, which is the characterization of Riemannian Greeks.

Remark 2.23 (Calibration weights). The theorem is no more true already in the case in which one adds calibration weights $w_i \in \mathbb{R}^+$:

$$f_i(\boldsymbol{\theta}, \mathbf{y}) = w_i(\phi_i(\boldsymbol{\theta}) - y_i). \quad (2.19)$$

Indeed, denoting by $\mathbf{W}$ the diagonal matrix with diagonal elements $w_1, \dots, w_n$, equation (2.17) becomes

$$\begin{aligned} (\text{grad } \tilde{P})^\top &= -\bar{\boldsymbol{\theta}} \cdot ((\mathbf{W}\mathbf{D}\phi)^\top (\mathbf{W}\mathbf{D}\phi))^{-1} (\mathbf{W}\mathbf{D}\phi)^\top (-\mathbf{W}) \\ &= \bar{\boldsymbol{\theta}} \cdot ((\mathbf{W}\mathbf{D}\phi)^\top (\mathbf{W}\mathbf{D}\phi))^{-1} (\mathbf{D}\phi)^\top \mathbf{W}^2, \end{aligned}$$

in which one sees that varying $\bar{\theta}_1, \dots, \bar{\theta}_m$, the transpose of the right hand side spans all values of $\text{Image}(\mathbf{W}^2\mathbf{D}\phi)$ which is distinct from $\text{Image}(\mathbf{D}\phi) = T_{\mathbf{y}}Y$ in general. In particular, there exists a choice of $\bar{\theta}_1, \dots, \bar{\theta}_m$ such that $\text{grad } \tilde{P}$ does not belong to $T_{\mathbf{y}}Y$, and hence cannot be the minimum-norm solution of (G).

The point is that for general $\mathbf{f}$, calibration is not projection with respect to the Euclidean distance any more, because the latter is incoherent with the “nearness ordering” induced by the object function. Hence we get the feeling that we may have chosen the wrong metric to draw the immersion of Y in $\mathbb{R}^n$. The intuition is right, but misses a little complication: we are not free to change the scalar product $\langle \mathbf{g}, \delta \mathbf{y} \rangle$ in the treatment above, because it arises just from a convenient rewriting of the right hand side of the fundamental equation (G), which equation is purely differential-geometric and does not involve any Riemannian metric.

However, we know from Remark 2.11 that by a change of coordinates we can alter the definition of the minimized norm. In the case of (2.19), there is a rescaling which restores the coincidence of the least-squares objective function $\sum_i w_i^2 (\phi_i(\boldsymbol{\theta}) - y_i)^2$ with the squared distance between the model and market point, and hence the coincidence of Riemannian and bestfit Greeks: just define $y'_i = w_i y_i$. This has a clear financial meaning, since we can reinterpret f_i as the error in calibrating to the price y'_i of a modified instrument with notional multiplied by w_i . Since we have redefined the instrument and the quote to incorporate w_i , the new “calibration instruments” contribute equally to the least-squares objective function, and we are back to the unweighted case of Theorem 2.22.

A final comment. We have seen that one cannot incorporate the effects of calibration weights in the Greeks equation, but we managed nonetheless to recover them with a careful choice of the minimized norm. This is not surprising, since it is unfair to ask (G) to take care directly of numbers which play a role only in configurations which the model excludes. In this respect, the strength of (G) is its weakness, i.e. its indeterminacy, which shapes a well-defined freedom to weight differently the various instruments used in hedging. Here is where market wisdom may play its crucial role in supplementing the model in issues it cannot capture. We will elaborate on this in Section 2.8.

2.7.2 If $\mathbf{y} \notin Y$

In this case, strictly speaking, our approach looks inapplicable, because function P is not even defined in $\mathbf{y}$. One may be tempted to conclude that the only possibility left is to differentiate $\tilde{P}$; however, the Greeks computed this way would be highly influenced by the bestfit calibration, i.e. an artificial relation between model parameters and values of $\mathbf{y}$ which the model itself excludes (empirical evidence presented in Section 2.9). The significance of the results would be therefore at very least questionable.

There is an alternative: differentiating P in a point $\mathbf{y}' \in Y$ close to the $\mathbf{y}$ originally given. In the simple case of Section 2.4, the natural choice would be $\mathbf{y}' = \phi(\theta)$, i.e. calculating the Greeks with respect to theoretical market quotes instead of observed ones: which is already common as a rule of thumb to stabilize the computation of sensitivities. After all, when one calibrates inexactly, and then uses the calibrated model to compute a price, the implicit assumption is that the “real” value of market instruments is $\mathbf{y}' = \phi(\theta)$ (the value we would get repricing the instruments themselves), while the fact that the observed values are $\mathbf{y} \neq \mathbf{y}'$ is neglected “noise”. In this interpretation, it makes sense to discard this noise also in the determination of Greeks.

A conclusive remark. In formulas (2.16), the computation of the second order derivatives is generally non-trivial and computationally costly. The advice of Henrard (2013) is to neglect such terms, on the ground that if calibration is good then such second derivatives are multiplied by f_i which is almost zero; indeed, the approximation becomes exact precisely if $\mathbf{y} \in Y$. We might say that the recipe of Henrard (2013) is to “compute an approximate derivative in the exact point”, while ours is to “compute the exact derivative in an approximate point”. Both approximations are increasingly good according to the accuracy of the calibration, but the second one allows to detect very clearly the financial information we are neglecting: it is the information in the calibration error. One cannot expect in general that one algorithm is more precise than the other, as the following theorem shows.

Theorem 2.24 (Meaning of the no-Hessian approximation). *Consider the simple case in which $f_i(\theta, \mathbf{y}) = \phi_i(\theta) - y_i$ with the hypotheses of Theorem 2.12. Then the following algorithms give the same result:*

1. *Substituting $\mathbf{y}$ with $\mathbf{y}' = \phi(\theta)$ and then computing the Riemannian Greek from (G).*
2. *Neglecting second derivatives in (2.16) in computing $\text{grad } \tilde{P}$ from (2.17).*

Proof. It is sufficient to remark that in the second algorithm, once neglected the second-derivative terms, equation (2.17) becomes exactly (2.18), which depends only on θ but no more on $\mathbf{y}$. Hence the thesis follows immediately by Theorem 2.22. $\square$

Example (conclusion). Let us come back one last time to our example

$$\phi_1(\theta, \mathbf{y}) = \cos \theta - y_1, \quad \phi_2(\theta, \mathbf{y}) = \sin \theta - y_2, \quad p(\theta) = \sin^2 \theta.$$

One readily sees that given any $\mathbf{y} \in \mathbb{R}^2 \setminus \{(0,0)\}$, the best θ , i.e. $\tilde{c}(\mathbf{y})$, is the one defined by direction $\frac{\mathbf{y}}{\|\mathbf{y}\|}$. Therefore

$$\tilde{P}(\mathbf{y}) = \sin^2 \tilde{c}(\mathbf{y}) = \frac{y_2^2}{y_1^2 + y_2^2}$$

is a 0-homogeneous function, hence its gradient is a -1 -homogeneous function. As a consequence, substituting $\mathbf{y}$ with $\mathbf{y}' = \phi(\theta) = \mathbf{y} / \|\mathbf{y}\|$ (which, as seen above, is equivalent to applying the approximation in Henrard (2013)), produces a result differing

from $\text{grad } \tilde{P}(\mathbf{y})$ by a factor $\|\mathbf{y}\|$. In particular, the relative error is $|\|\mathbf{y}\| - 1|$, i.e. the distance of $\mathbf{y}$ from Y .

To sum up, if one really wants $\text{grad } \tilde{P}$, the accuracy of the approximation is measured by the distance from the set of admissible quotes Y : our framework makes this result look extremely natural. On the other hand, now we know that the algorithm computes exactly $\text{grad } \tilde{P}(\mathbf{y}')$, which under a different point of view might be even more appropriate than $\text{grad } \tilde{P}(\mathbf{y})$ itself.

2.8 Considerations on the choice of the representative

As we saw, all that mathematics and the model can give us is an undetermined constraint for $\mathbf{g}$. The choice of the particular representative depends necessarily on the context and on extra-model considerations, and one cannot expect a recipe to be good on all occasions, as the following example shows.

Example. Consider a market with a single equity underlying S_t , quoting two at-the-money call options with maturity 5 and 10 years, by their implied volatilities Σ_5 and Σ_{10} . Suppose that the following log-normal model is a satisfactory description of the dynamics (interest rates are null for simplicity):

$$dS_t = \sigma(t)S_t dW_t, \quad (W_t) \text{ Brownian motion, } \sigma(t) = \sigma_5 I_{[0,5)}(t) + \sigma_{10} I_{[5,10)}(t).$$

If $\sigma_5 \approx \sigma_{10}$, one can imagine that for the purpose of pricing the following approximate model has been considered acceptable:

$$dS_t = \bar{\sigma} S_t dW_t, \quad \bar{\sigma} \in \mathbb{R}. \quad (2.20)$$

The value of $\bar{\sigma}$ will be obtained calibrating to the two options; the exact number will depend on the concrete form chosen for the objective function, but it will be for sure an increasing function of Σ_5 and Σ_{10} , for instance $\bar{\sigma} = (\Sigma_5 + \Sigma_{10})/2$.

Let us consider now the pricing of two different products:

1. The 5-years-maturity option used in calibration.
2. A forward-start option with start date in 5 years and maturing in 10 years.

The approximate model gives the same value to the two options, given by a Black formula:

$$p(\bar{\sigma}) = \text{BS}(S, S, \bar{\sigma}, 5);$$

In particular, denoting by ν the “Vega” of the formula (derivative with respect to volatility), both the Riemannian Greeks and the bestfit Greeks will be something like:

$$g_5 = g_{10} = \frac{1}{2}\nu.$$

However, with the exact model, the best answer would be the following:

1. For the spot-start option, $g_5 = \nu$ and $g_{10} = 0$. In this circumstance, a wisely-chosen economical Greek would have been perfect.
2. For the forward-start option, the exact price would have been

$$\text{BS}(S, S, \sqrt{2\Sigma_{10}^2 - \Sigma_5^2}, 5)$$

whose derivatives are

$$g_5 = \frac{-\nu \Sigma_5}{\sqrt{2\Sigma_{10}^2 - \Sigma_5^2}} \approx -\nu, \quad g_{10} = \frac{2\nu \Sigma_{10}}{\sqrt{2\Sigma_{10}^2 - \Sigma_5^2}} \approx 2\nu.$$

Indeed, if at the moment of calibration $\Sigma_5 = \Sigma_{10}$ holds exactly, then $(g_5, g_{10}) = (-\nu, 2\nu)$ is a valid Greek in our framework; but we cannot expect to single out this representation, or even only the negative sign of g_5 , using only (2.20) and its calibration.

Nonetheless, our framework clearly identifies the alternatives, and provides a vocabulary to analyse their pros and cons. For instance, opinions like the following become natural:

- In the market I trade into, using many hedging instruments is costly and/or impractical: I will go for the economical Greeks.
- Some hedging instruments are not liquid: I will spend some degrees of freedom to put to zero the corresponding Greeks, and will then perform some Riemannian minimization on the remaining components.
- I have extra-model views hinting at which pillars are more relevant for the position I have to hedge: I will help the sensitivities engine asking for economical Greeks concentrated on those pillars.
- I want a synthetic scalar indicator to know at each moment “how much my portfolio is imbalanced”; a quadratic form in my Greeks will do. Of course I will have to compute the Greeks in the corresponding generalized Riemannian form, to be sure that I am measuring the net imbalance which cannot be cancelled using the degrees of freedom in the definition of $\mathbf{g}$.
- I would like to weight differently the various hedging instruments in my portfolio, but I do not want to touch the calibration weights because of the side-effects on calibrated parameters. Never mind, I will tweak the Riemannian metric minimized in the computation of my Greeks.
- To monitor whether market risks are correctly quantified, I verify periodically the P&L Explain, i.e. whether the profit or the loss is well explained by the first order expansion $\mathbf{g} \cdot \Delta \mathbf{y}$. My profit or loss is defined as $\Delta \tilde{P}$ (recalibrate-then-reprice), hence by definition the most performing $\mathbf{g}$ in this sense will be the bestfit Greeks $\text{grad}(\tilde{P})$.

One can note that the above needs, leading to different choices of the representative, are not necessarily alternative, since they refer to different moments in the trading activity. As an example, in principle one may:

- Use the Riemannian Greeks to decide *whether* to intervene on the hedging portfolio.
- Once decided to re hedge, use economical Greeks to minimize the cost of the transaction.
- Use anyway the bestfit Greeks in the P&L explain.

2.9 Numerical experiments

We have empirically compared the different methodologies presented above, in the following model setting for Euro-denominated interest rates:

- The stochastic discount rate is given by the EONIA rate, whose forwards seen at the evaluation date are extracted from the quotation of Overnight Indexed Swaps (OIS). The EONIA rate is modelled by a Hull&White process $r(t)$:

$$dr(t) = k(\theta(t) - r(t))dt + \sigma dW_t,$$

in which $\theta(t)$ is a deterministic function, k and σ are real numbers, and W_t is a standard Brownian motion. Parameter θ is time-dependent so that, once fixed the parameters k and σ , there exists a unique $\theta(\cdot)$ able to fit exactly the yield curve at the evaluation date.

- The interest rate EURIBOR 6M is itself deduced from a short-rate model, potentially different from the one for EONIA (multicurve pricing, see e.g. Mercurio, 2010), but with the hypothesis that such short rate is given by a deterministic spread $\phi(t)$ over EONIA. The spread is time-dependent and is chosen to fit the EURIBOR 6M curve at the evaluation date.

As we said, since we desire to fit exactly the rate curves at the evaluation date, the quantities $\theta(t)$ and $\phi(t)$ are not free parameters, because they are determined once chosen k and σ , which are therefore the only parameters at our disposal to fit the volatility structure of interest rates. The market information on this volatility is shipped by swaption quotes, which are varied by expiry of the option, maturity of the underlying swap, and strike. Since we have only two free parameters, the calibration will consider only a limited number of at-the-money swaptions, selected among the most actively traded.

With the calibrated model, we then priced a non quoted option expiring nearly in 4 years and a half, to enter in 5-years-long swap. This product is deliberately similar but not identical to one of the instruments used in calibration, namely the 5x5 swaption, so that it is easy to have an intuition (right or wrong) on how the Greeks should look like.

In the graphics of this section:

- “Riemannian” denotes the Riemannian Greek;
- “Economical” denotes the economical Greek zeroing the sensitivity to all swaptions except the 5x5 and 5x10 ones;
- “Theoretical bestfit” denotes the bestfit Greek *computed substituting market data with the theoretical provisions of the calibrated model*.
- “No-hessian bestfit” denotes the bestfit Greek *computed ignoring Hessians as suggested in Henrard (2013)*;

We proceed commenting the results obtained with market data as of 12 October 2015, for different choices of the objective function $\mathbf{f}$. In the following subsection, we start with the simple and well-understood case

$$f_i = (\text{Model price} - \text{Quoted price}) \text{ of the } i\text{-th market swaption}, \quad (2.21)$$

in which case we know that many approaches collapse to the same result, and then see what happens when each f_i is multiplied by a weight w_i . In the subsequent subsection, instead, we consider a more nonlinear but equally reasonable form for f_i , to see how this impacts the various representatives for $\mathbf{g}$ and how different among themselves they can become.

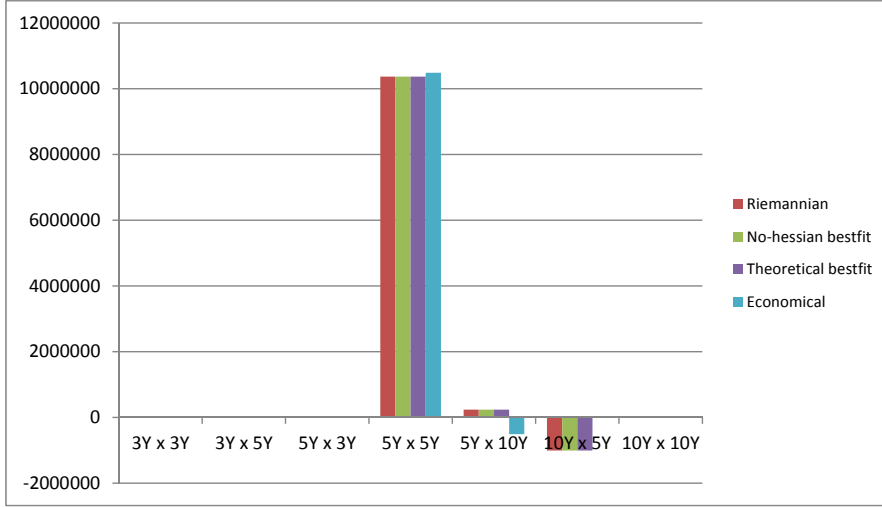


Figure 2.3: Sensitivity of the NPV of a non quoted swaption, with an object function which is linear in market prices. Market data: premia of three quoted swaptions. Equally-weighted calibration.

2.9.1 Simple or weighted-simple object function

In the experiments described in this subsection, each component of $\mathbf{f}$ will have the form (2.21) apart possibly for a multiplicative weight w_i which can be a different nonnegative number for each i .

Firstly, we performed the experiment calibrating the model only to the three equally-weighted swaptions 5x5, 5x10 and 10x5, so that the indeterminacy affecting the Greeks is only of dimension 1. Figure 2.3 shows that all methods give similar results: as one would expect, they attribute almost all the sensitivity to the 5x5 pillar; since the product is not exactly the 5x5 swaption, it is not surprising the appearance of some residual sensitivity on the two other pillars, although they involve quite longer maturities. Note that the Riemannian Greeks and the two bestfit variants yield exactly the same result, as foreseen by Theorem 2.24.

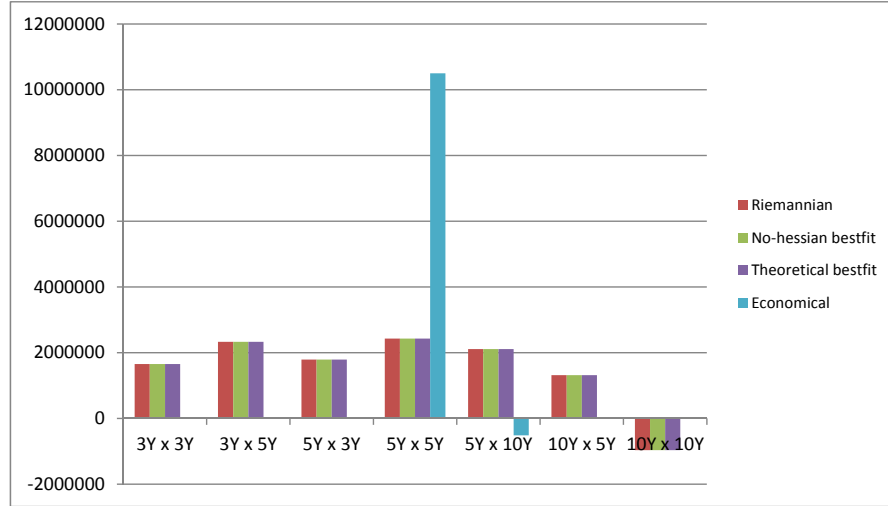
Then, we enlarged the perimeter of calibration instruments to seven swaptions. To study the effect of calibration weights on results, we performed the experiment both with uniform weights, and overweighting the 5x5 swaption by a factor five and the 5x10 swaption by a factor two. Results are in Figure 2.4, and they show that:

- With the obvious exception of the economical Greek, when weights are uniform all methodologies spread the sensitivity widely on all pillars (Figure 2.4a).
- The change in calibration weights changes radically the bestfit Greeks, while it impacts only marginally the economical and Riemannian Greeks, which are more intrinsic (Figure 2.4b).
- The two bestfit Greeks coincide perfectly.

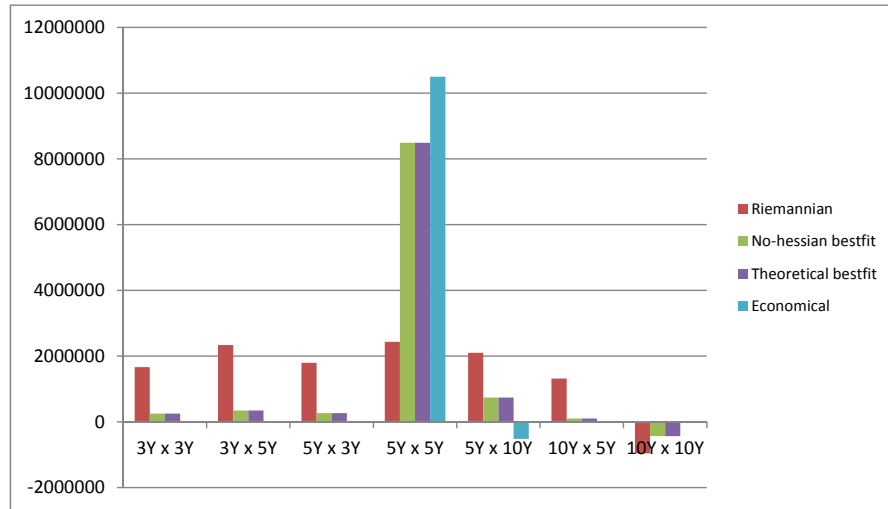
2.9.2 General object function

Numerical considerations often suggest alternative forms for the object functions like

$$f_i = w_i \left(\frac{\text{Model price}}{\text{Quoted price}} - 1 \right) \text{ of the } i\text{-th market swaption, } w_i \in \mathbb{R}^+. \quad (2.22)$$



(a) Uniform calibration weights.



(b) Calibration weights: 5.0 for the 5x5 swaption, 2.0 for the 5x10 swaption, 1.0 for the remaining swaptions.

Figure 2.4: Sensitivity of the NPV of a non quoted swaption, with an object function which is linear in market prices, and with two different choices of calibration weights. Market data: premia of seven quoted swaptions.

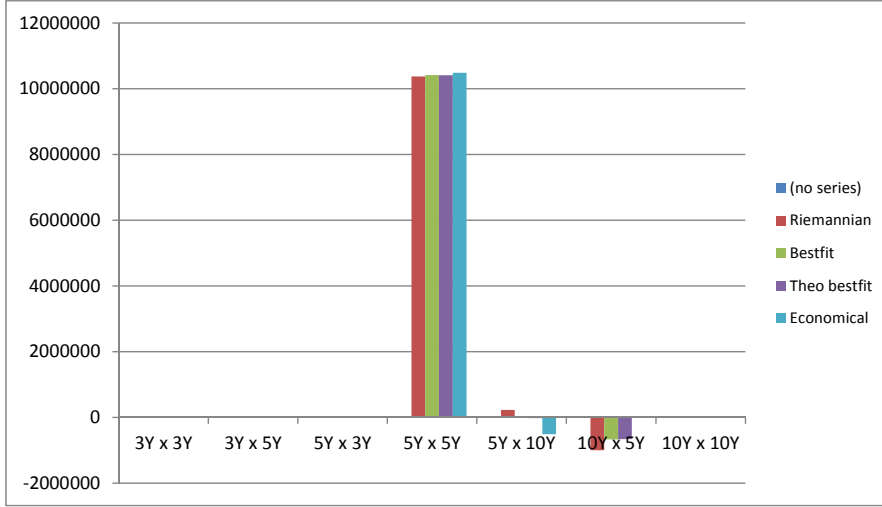


Figure 2.5: Sensitivity of the NPV of a non quoted swaption, with an object function which is nonlinear in market prices. Market data: premia of three quoted swaptions. Equally-weighted calibration.

Therefore, we repeated our tests with this kind of object function, and we now present the main differences which emerged.

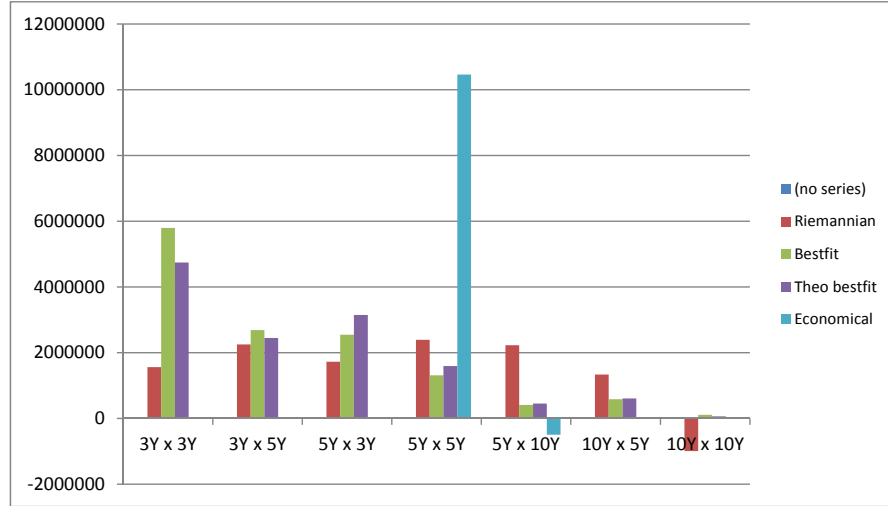
The experiment calibrating the model only to three equally-weighted swaptions presents no surprise. Figure 2.5 shows that all methods give similar results, again with most of the sensitivity on the 5x5 pillar, and some unavoidable remainders on the two other pillars. However, note that this time each algorithm has a slightly different outcome.

The differences widen if one calibrates to seven swaptions. Results are in Figure 2.6, which allows the following remarks:

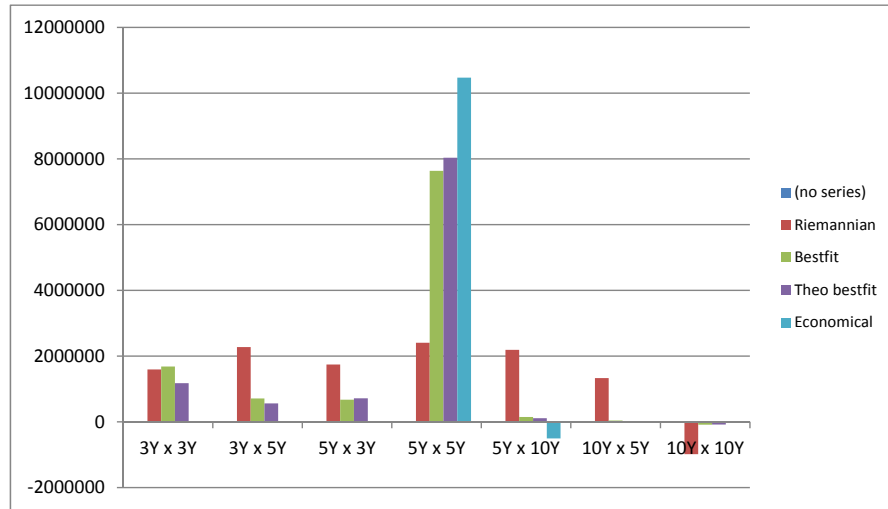
- Again, when weights are uniform all the non-economical methodologies feature a dispersion of significant sensitivities on all pillars (Figure 2.6a). Moreover, this time the bestfit Greeks have a potentially improper peak on the shortest pillar 3x3.
- The change of object function affects considerably the bestfit Greeks, while the economical and Riemannian Greeks maintain their profile. This is the non-parametric counterpart of the effect of calibration weights, already remarked in the previous subsection and also evident in Figure 2.6b.
- The two bestfit Greeks are similar, but the differences are not always completely negligible (especially with uniform weights).

2.10 Conclusions and further research

Whether on a basis of a full theoretical justification or not, real-life hedging requires the Greeks of the price function. This chapter shows the power and expressiveness of the language of differential geometry in describing the problem of computing them. Thanks to this formalism, we saw how in presence of calibrations one cannot give an univocal solution to the problem, and we characterized completely the range of possible solutions in terms of simple linear systems.



(a) Uniform calibration weights.



(b) Calibration weights: 5.0 for the 5x5 swaption, 2.0 for the 5x10 swaption, 1.0 for the remaining swaptions.

Figure 2.6: Sensitivity of the NPV of a non quoted swaption, with an object function which is nonlinear in market prices, and with two different choices of calibration weights. Market data: premia of seven quoted swaptions.

We then showed how the theoretical framework can help greatly in formulating practical considerations and in taking rational decisions when devising realistic hedging strategies. Different representations for the set of Greeks can be useful in different contexts, without necessarily excluding one another.

We believe that this work could be a good starting point for anybody looking for the “best” Greeks, in some sense to be discussed and defined case by case. We note in particular that the continuum of generalized Riemannian Greeks, which minimize a generic quadratic penalty function, may be rich enough for a rigorous treatment of liquidity conditions, transaction costs, P&L volatility and more generally trader’s preferences. Even more ambitious would be the study and optimization of a dynamical hedging portfolio adjusted with the passage of time, exploiting the degrees of freedom pointed out by this treatment.

Finally, we remark that all the variants would be theoretically equivalent provided that the chosen parametrization of the dynamics is sufficiently rich to represent satisfactorily each possible configuration of the market. This could suggest a way to test if a chosen model is flexible enough; the resulting validation approach would be more coherent with practice, since it would measure the ability to follow quote movements allowing recalibration of parameters, rather than in the school hypothesis that they are universal constants.

Part II

Fast computation of multiple sensitivities

Chapter 3

Fast first order sensitivities for discontinuous payoffs

In this chapter, we present a general technique to compute the sensitivities of the Monte Carlo prices of possibly discontinuous financial products. It is a natural extension of the pathwise adjoints method, which would require an almost-surely differentiable payoff; the efficiency of the latter method when many sensitivities must be calculated is preserved. We show empirically that the new algorithm is competitive in terms of accuracy and execution time when compared to benchmarks obtained by smoothing of the payoff, which benchmarks are biased and require a non-obvious tuning of their parameters.¹

3.1 Introduction

As we have seen in the first part of this thesis, in the theory and practice of the management of financial portfolios, a crucial role is played by the valuation not only of the present value of the claims, but also of their sensitivity to market movements, in order to correctly quantify and hedge risks. The centrality of such sensitivities or Greeks is confirmed by their growing use in the industry to define standard risk-sensitive metrics. For instance, the new standard for contracts proposed by the International Swaps and Derivatives Association (2013, 2016) uses sensitivities to compute initial margins; while regulators have introduced a sensitivity-based approach in their directives on capital requirements (Basel Committee on Banking Supervision, 2013).

On the algorithmic side, the main challenge is the large number of sensitivities to compute. Even a simple interest rate swap depends potentially on tens of market quotes, corresponding to the various pillars of the discounting and forwarding curves; the picture worsens if more risk drivers are involved, for instance when the computation must be performed at portfolio level, as is the case for the determination of non-linear fair-value adjustments like the CVA (Basel Committee on Banking Supervision, 2015).

The recent popularity of this kind of metrics has given new impulse and relevance to the problem of efficiently computing sensitivities. In fact, the simplest approach is to compute them numerically, moving each market datum individually and repeating for each of these movements the whole pricing algorithm. This is an extremely expensive

¹A slightly different version of this chapter has been published in Daluiso and Facchinetti (2018).

procedure in terms of computational time, especially when the price is itself the result of a costly Monte Carlo simulation.

In this context, a very promising technique is the so-called adjoint differentiation, based on the analytical computation of a sequence of intermediate derivatives, which gives the full gradient of a scalar function in at most four times the computational time of the implementation of the function alone (Griewank and Walther, 2008). However, the direct application of this methodology to a Monte Carlo pricing is possible only if the payoff is continuous and almost surely differentiable, so that the sensitivities can be computed path-by-path (Glasserman, 2004). This is not the case for many financial products, which have digital or barrier features. For such products one has to substitute or complement these algorithmic differentiation (AD) techniques, exploiting the nature of the problem.

Some AD-based solutions have been proposed in the 2000s, but they are not sufficiently generic for large-scale application. In particular, the likelihood-ratio method (Glasserman, 2004) can compute the sensitivity to a parameter only if it influences the price exclusively through the probability density of the underlying; moreover, if the density at maturity is unknown and must be approximated via a simulation scheme, the statistical uncertainty of the Greek becomes very large. The variant called vibrato (Giles, 2007) solves in most cases the variance issue, but still does not handle direct dependencies of the payoff on parameters: in particular, neither of these methods is applicable as-is to the very common setting of factorial models, which express many underlyings as functions of a relatively parsimonious set of risk factors; or to the computation of sensitivities to deterministic rate curves.

If one drops for a moment the desire to compute high-dimensional gradients quickly and looks for estimators of single Greeks, then more material can be found in the literature. Perhaps the earliest results have been obtained porting to finance the ideas of perturbation analysis, see e.g. Fu and Hu (1997) and references therein. Turning to more field-specific recipes, we mention the proposals based on Malliavin calculus (Malliavin and Thalmaier, 2006; Fournié et al., 1999, 2001; Benhamou, 2003), which suffer from similar limitations as those of likelihood ratio. A higher level of generality has been recently achieved in the fairly abstract approach of Feng and Liu (2016), whose concrete instantiations require each time the manual or numerical computation of at least one integral depending on the model and the payoff; one of their formulations essentially reduces to the earlier findings of Lyuu and Teng (2011). The latter are very related to the formulas of our Section 3.3, even though the hypotheses and final result do not exactly overlap.

In none of the papers cited in the previous paragraph, the algorithmic aspects or computational efficiency are discussed, especially in relation to scalability when many sensitivities must be computed. To our knowledge, the only work addressing this kind of issues in a sufficiently generic setting is Chan and Joshi (2015); here we provide an alternative which is different in both spirit and implementation, whose closer relative is arguably the technique developed in Capriotti et al. (2011) for the specific application of counterparty credit risk management.

This chapter offers a direct generalization of the pathwise method. The key observation is that if the payoff has a discontinuity, then it contributes to the derivative as a Dirac delta. The result is a closed-form unbiased estimator, which can be computed with adjoint differentiation, and which reduces to the classical pathwise estimator if the payoff is continuous. The general idea is developed in full detail in the case of foremost practical relevance, i.e. a Monte Carlo with diffusive risk factors: in such context, a clever organization of the simulation algorithm allows a drastic reduction

of the computational overhead of the terms generated by the discontinuity relative to the original pricing time.

We conclude with some numerical examples. As a benchmark for the result, for the uncertainty and for the execution times, in each example the sensitivities are also calculated pathwise after smoothing of the payoff. This touchstone is chosen for its generality and popularity; even though it involves a delicate tradeoff between the bias introduced by the approximation and the Monte Carlo uncertainty of the Greek, especially in presence of multiple discontinuities, as we will see. The same kind of qualitative remarks would apply to more sophisticated but still smoothing-based biased approaches like Liu and Hong (2011).

The rest of the chapter is organized as follows. Section 3.2 introduces the key idea using the stylized example of a binary call; Section 3.3 generalizes to arbitrary payoffs. Section 3.4 briefly introduces adjoint algorithmic differentiation, which is used to formulate a fast algorithm to compute the formulas of the preceding sections; an analysis of the complexity of the algorithm then suggests a variant in which the largest possible number of intermediate results is reused from the pricing phase. Section 3.5 applies all of the above in the case of a Monte Carlo with diffusive risk factors, showing that in such setting the computations which can be saved are indeed many. Section 3.6 contains several numerical experiments, and Section 3.7 summarizes the main contributions and hints at possible applications. Finally, Appendix A.2 provides technical hypotheses under which we can prove with full rigour the main formulas, which in the body of the chapter are derived sacrificing details to underline the ideas.

3.2 The digital option

In this section we will use the pricing of a binary option as an introductory example. The derivation is intentionally sometimes heuristic, while detailed hypotheses and proofs are available in Appendix A.2.

Let us consider the price of a digital option:

$$P = \mathbb{E}[I_{X(T)>0}], \quad X(T) = f(\boldsymbol{\theta}, Z_1, \dots, Z_d) =: f(\boldsymbol{\theta}, \mathbf{Z}), \quad (3.1)$$

where the underlying $X(T)$ is computed on simulated scenarios, and can therefore be expressed as a function f of $Z_i \sim \mathcal{N}(0, 1)$ independent and identically distributed and of model parameters $\boldsymbol{\theta}$. Our goal is the computation of $\partial P / \partial \boldsymbol{\theta}$.

Let us rewrite the expected value as an integral and move the derivative operator inside:

$$\begin{aligned} \frac{\partial P}{\partial \boldsymbol{\theta}} &= \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}[I_{f(\boldsymbol{\theta}, Z_1, \dots, Z_d) > 0}] \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^{d-1}} \int_{-\infty}^{+\infty} I_{f(\boldsymbol{\theta}, z_1, \dots, z_d) > 0} e^{-z_k^2/2} dz_k \prod_{i \neq k} e^{-z_i^2/2} dz_i \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^{d-1}} \int_{-\infty}^{+\infty} \left[\delta(f) \frac{\partial f}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, z_1, \dots, z_d) e^{-z_k^2/2} dz_k \prod_{i \neq k} e^{-z_i^2/2} dz_i, \end{aligned} \quad (3.2)$$

where in the last passage we used formally the chain rule and the fact that the derivative of an indicator function is a Dirac delta:

$$\frac{\partial}{\partial \boldsymbol{\theta}} I_{f(\boldsymbol{\theta}, z_1, \dots, z_d) > 0} = \frac{\partial}{\partial f} I_{f > 0} \cdot \frac{\partial f}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, z_1, \dots, z_d) = \left[\delta(f) \frac{\partial f}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, z_1, \dots, z_d). \quad (3.3)$$

To proceed, we introduce some notations: given a vector $\mathbf{v} \in \mathbb{R}^d$ (in boldface) and $k \in \{1, \dots, d\}$, we will denote by $\mathbf{v}_{-k} \in \mathbb{R}^{d-1}$ the vector obtained from $\mathbf{v}$ removing its k -th component, and with $\mathbf{v}(v_k = x)$ the vector obtained from $\mathbf{v}$ substituting the k -th component with the value x . Finally, we will denote by $\mathbf{I}_{\mathbf{v} > \mathbf{0}}$ the vector $(I_{v_1 > 0}, \dots, I_{v_d > 0})$.

Now we can formulate the following hypothesis:

Hypothesis 3.1. f is differentiable with respect to $\boldsymbol{\theta}$ and z_k ; for every $\mathbf{Z}_{-k}$, the function $z_k \mapsto f(\boldsymbol{\theta}, \mathbf{Z}(Z_k = z_k))$ is zero only for z_k in a finite set $A_k(\boldsymbol{\theta}, \mathbf{Z}_{-k})$, and its derivative in such points is not vanishing.

Thanks to the above hypothesis, for each $\mathbf{Z}_{-k}$ fixed we can use a well known representation of the composition of the delta distribution with a function ϕ of a single real variable x (Schwartz, 1966):

$$\delta(\phi(x)) = \sum_{\bar{x}: \phi(\bar{x})=0} \left| \frac{\partial \phi}{\partial x}(\bar{x}) \right|^{-1} \delta(x - \bar{x}), \quad (3.4)$$

which is valid under suitable conditions on ϕ . Hence, we can continue integrating the Dirac delta:

$$\begin{aligned} \frac{\partial P}{\partial \boldsymbol{\theta}} &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^{d-1}} \sum_{\zeta_k \in A_k(\boldsymbol{\theta}, \mathbf{Z}_{-k})} \left[\left| \frac{\partial f}{\partial z_k} \right|^{-1} \frac{\partial f}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_k = \zeta_k)) e^{-\zeta_k^2/2} dz_k \prod_{i \neq k} e^{-z_i^2/2} dz_i \\ &= \mathbb{E} \left[\sum_{\zeta_k \in A_k(\boldsymbol{\theta}, \mathbf{Z}_{-k})} \frac{e^{-\zeta_k^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f}{\partial z_k} \right|^{-1} \frac{\partial f}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_k = \zeta_k)) \right]. \end{aligned} \quad (3.5)$$

3.3 Arbitrary combinations of indicator functions

The reasoning easily generalizes to

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E} [I_{f(\boldsymbol{\theta}, \mathbf{Z}) > 0} \cdot g(\mathbf{Z})] &= \\ \mathbb{E} \left[\sum_{\zeta_k \in A_k(\boldsymbol{\theta}, \mathbf{Z}_{-k})} \frac{e^{-\zeta_k^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f}{\partial z_k} \right|^{-1} \frac{\partial f}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_k = \zeta_k)) \cdot g(\mathbf{Z}(Z_k = \zeta_k)) \right]; \end{aligned} \quad (3.6)$$

now if h functions f_i for $i = 1, \dots, h$ are given, each satisfying Hypothesis 3.1 for a set of roots $A_{k(i)}^i$, and if g is a function of $(\boldsymbol{\theta}, \mathbf{Z})$ such that $\partial g / \partial \boldsymbol{\theta}$ exists, with an obvious argument based on the derivative of a composition we get:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E} \left[\prod_{i=1}^h I_{f_i(\boldsymbol{\theta}, \mathbf{Z}) > 0} \cdot g(\boldsymbol{\theta}, \mathbf{Z}) \right] &= \mathbb{E} \left[\prod_{i=1}^h I_{f_i(\boldsymbol{\theta}, \mathbf{Z}) > 0} \cdot \frac{\partial g}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{Z}) \right] + \\ &+ \sum_{i=1}^h \mathbb{E} \left[\sum_{\zeta \in A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})} \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \prod_{j \neq i} I_{f_j > 0} \cdot g \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \right]. \end{aligned} \quad (3.7)$$

(For a proper proof of (3.6) with the specification of a set of sufficient technical hypotheses, refer to Appendix A.2.)

Finally, with h and f_i as above, and denoting by $\mathbf{f}$ the vector valued function with components f_i , we consider any payoff of the form

$$P = \mathbb{E} [\phi (I_{f_1}(\boldsymbol{\theta}, \mathbf{Z}) > 0, \dots, I_{f_h}(\boldsymbol{\theta}, \mathbf{Z}) > 0, \boldsymbol{\theta}, \mathbf{Z})] = \mathbb{E} [\phi (\mathbf{I}_{\mathbf{f}}(\boldsymbol{\theta}, \mathbf{Z}) > \mathbf{0}, \boldsymbol{\theta}, \mathbf{Z})], \quad (3.8)$$

for arbitrary ϕ differentiable with respect to $\boldsymbol{\theta}$. The differentiation of such P can be recast to a set of exercises of the form (3.7), because

$$\phi (\mathbf{I}_{\mathbf{f}}(\boldsymbol{\theta}, \mathbf{Z}) > \mathbf{0}, \boldsymbol{\theta}, \mathbf{Z}) = \sum_{\mathbf{a} \in \{0,1\}^h} \phi (\mathbf{a}, \boldsymbol{\theta}, \mathbf{Z}) \prod_{i=1}^h I_{(-1)^{a_i+1} f_i}(\boldsymbol{\theta}, \mathbf{Z}) > 0. \quad (3.9)$$

as one can verify since the left hand and right hand sides trivially coincide on each of the sets $\cap_{i=1}^h \{(-1)^{a_i+1} f_i(\boldsymbol{\theta}, \mathbf{Z}) > 0\}$, which form a partition of $\mathbb{R}^d$ as $\mathbf{a} = (a_1, \dots, a_h)$ varies in $\{0,1\}^h$ (the sets $\{f_i = 0\}$ are negligible, since their sections are finite by hypothesis.)

To write the final result in a handy way, we introduce the following notation: for a function $\psi : \{0,1\}^h \rightarrow \mathbb{R}$ and $i = 1, \dots, h$, we define

$$\Delta^{(i)} \psi(\mathbf{a}) = \psi(\mathbf{a}(a_i = 1)) - \psi(\mathbf{a}(a_i = 0)). \quad (3.10)$$

Then

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E} [\phi (I_{f_1}(\boldsymbol{\theta}, \mathbf{Z}) > 0, \dots, I_{f_h}(\boldsymbol{\theta}, \mathbf{Z}) > 0, \boldsymbol{\theta}, \mathbf{Z})] = D_0 + \sum_{i=1}^h D_i, \quad (3.11)$$

where D_i are given by:

$$\begin{aligned} D_0 &= \mathbb{E} \left[\sum_{(a_1, \dots, a_h) \in \{0,1\}^h} \frac{\partial \phi}{\partial \boldsymbol{\theta}} (a_1, \dots, a_h, \boldsymbol{\theta}, \mathbf{Z}) \prod_{i=1}^h I_{(-1)^{a_i+1} f_i}(\boldsymbol{\theta}, \mathbf{Z}) > 0 \right] \\ &= \mathbb{E} \left[\frac{\partial \phi}{\partial \boldsymbol{\theta}} (I_{f_1}(\boldsymbol{\theta}, \mathbf{Z}) > 0, \dots, I_{f_h}(\boldsymbol{\theta}, \mathbf{Z}) > 0, \boldsymbol{\theta}, \mathbf{Z}) \right], \end{aligned} \quad (3.12)$$

$$\begin{aligned} D_i &= \mathbb{E} \left[\sum_{\zeta \in A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})} \left\{ \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \right. \right. \\ &\quad \times \sum_{a_i \in \{0,1\}} (-1)^{a_i+1} \sum_{\mathbf{a}_{-i} \in \{0,1\}^{h-1}} \prod_{j \neq i} I_{f_j}(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) > 0 \cdot \phi (\mathbf{a}, \boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \left. \left. \right\} \right] \\ &= \mathbb{E} \left[\sum_{\zeta \in A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})} \left\{ \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \right. \right. \\ &\quad \times \Delta^{(i)} \phi (\mathbf{I}_{\mathbf{f}}(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) > \mathbf{0}, \boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \left. \left. \right\} \right], \end{aligned} \quad (3.13)$$

the last equality coming from the fact that because of the product of indicator functions, the only addend of the inner summation on $\mathbf{a}_{-i}$ which does not vanish is the one with $\mathbf{a}_{-i} = (\mathbf{I}_{\mathbf{f} > \mathbf{0}})_{-i}$.

3.4 Algorithmic differentiation and preprocessing

This section analyses the computational cost of the formulas of the preceding sections. This will suggest a variant that can significantly reduce the execution time in some circumstances, which we will extensively use in the subsequent sections.

3.4.1 Classical algorithmic differentiation

This subsection briefly reviews the technique known as adjoint algorithmic differentiation (AAD), which we intend to exploit in our algorithms. The aim is only to setup a minimum vocabulary: interested readers can refer to Griewank and Walther (2008) for the general theory, and to Capriotti (2011); Capriotti and Giles (2012); Giles (2007) for an introduction to financial applications.

The idea of algorithmic differentiation is that if F is a function computed by a computer, then it is a composition of few types of elementary operations, either binary (additions, subtractions, multiplications, divisions) or unary (exponentials, logarithms, trigonometric functions), whose derivatives are well known: hence one can differentiate the full algorithm applying the chain rule. In particular, *adjoint* algorithmic differentiation prescribes to apply the chain rule from the last instruction to the first one, since this turns out to be computationally beneficial if the inputs of F are many and the outputs are few.

In practice, this means that the original algorithm, which in this context is called *forward sweep*, is followed by a *backward sweep* in which for each variable v in the computer program, a new variable $\bar{v}$ is allocated where derivatives are accumulated. If $F = (F_1, \dots, F_m)$ is a function of $\mathbf{x} = (x_1, \dots, x_n)$ and the backward sweep is initialized with some $\bar{F} = (\bar{F}_1, \dots, \bar{F}_m)$, then at the end of the backward sweep $\bar{\mathbf{x}}$ equals $\bar{F} \cdot DF$, where DF denotes the Jacobian of $\mathbf{x} \mapsto F$. In particular, if F is scalar-valued, initializing $\bar{F}$ to one will allow the computation of the gradient $\bar{\mathbf{x}} = (\nabla F)^\top$.

It is not hard to prove (Griewank and Walther, 2008) that the backward sweep has a theoretical computational cost of at most three times that of the original implementation, regardless of the number n of inputs. This property is very attractive in many financial applications, in which n is very large (many tens or hundreds of market data) while the number of outputs may be as small as one (the price).

Specifically, in the Monte Carlo estimation of $P = E[\pi]$, if the discounted payout π as a function of $\boldsymbol{\theta}$ is continuous and almost surely differentiable, then under weak hypotheses the price satisfies (Glasserman, 2004)

$$\frac{\partial P}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}} E[\pi] = E \left[\frac{\partial \pi}{\partial \boldsymbol{\theta}} \right], \quad (3.14)$$

and therefore $\partial P / \partial \boldsymbol{\theta}$ itself can be estimated as a Monte Carlo average of the samples of $\partial \pi / \partial \boldsymbol{\theta}$. This is very fit for the application of algorithmic differentiation, which gives $\partial \pi / \partial \boldsymbol{\theta}$ on each scenario quickly. The popularity of adjoints in finance is mainly linked to pricing problems of this sort.

However, a direct application of algorithmic differentiation is not possible if the payoff π is a discontinuous function of $\boldsymbol{\theta}$. Fortunately, in real payoffs, the discontinuities are always caused by conditional clauses (the flows change depending on future events), so that the most general payoff can always be expressed in the form (3.8). Indeed, the requirement that the f_i are differentiable is not restrictive in practice, since they always are at least piecewise, so that adding some more indicator functions

to the representation one can always recast the payoff into the form (3.8): for instance

$$I_{X_1 \vee X_2 > K} = I_{X_1 - X_2 > 0} I_{X_1 - K > 0} + I_{X_2 - X_1 > 0} I_{X_2 - K > 0}. \quad (3.15)$$

Hence we can hope to use the results of the previous section. In the following we will discuss how they can be efficiently combined with the ideas and algorithms of this subsection.

3.4.2 Computational graphs

To ease the subsequent exposition, this subsection introduces some graphical conventions which will allow the description of the valuation procedure for a payoff with a diagram.

In Section 3.4.1 we saw that every algorithm can be represented as a composition of elementary operations. This composition can be described by a computational graph, i.e. a set of vertices (or nodes) linked by oriented connections (or arcs) representing direct dependencies. Depending on contexts and on the desired granularity in the representation, arcs will represent elementary arithmetic operations or sometimes compositions of them. If the border of the node is doubled, we mean that the quantity may be a vector.

We will graphically distinguish the connections into three categories:

1. Smooth connections : they represent dependencies which are a composition of deterministic differentiable functions.
2. Random connections : differentiable operations involving the generation of one or more random variables.
3. Heaviside connections : non-differentiable dependencies consisting in the application of the Heaviside step function $H(x) = I_{x > 0}$. The result node will be drawn in triangular shape.

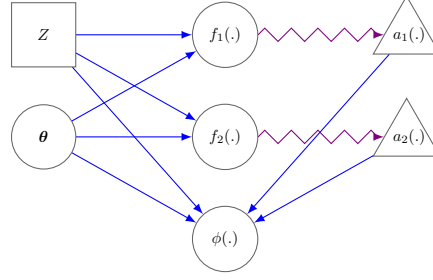
3.4.3 Algorithmic differentiation with indicator functions

Consider the pricing problem in its general form (3.8). Suppose that the price P is computed as a sample average of pseudo-random realizations of the discounted payoff ϕ . The typical implementation for the value of such realizations is schematically the following:

Algorithm 3.2. [Forward sweep] For each Monte Carlo scenario:

1. Generate random Gaussian numbers $Z_1, \dots, Z_d$.
2. For $i = 1, \dots, h$: compute $f_i(\boldsymbol{\theta}, \mathbf{Z})$.
3. Compute $\phi(I_{f(\boldsymbol{\theta}, \mathbf{Z}) > 0}, \boldsymbol{\theta}, \mathbf{Z})$.

In the example case $h = 2$ and $d = 1$, the computational graph would be:



A direct implementation of (3.11) would be the following:

Algorithm 3.3. [Backward sweep] For each Monte Carlo scenario, after the forward sweep:

2. For $i = 1, \dots, h$:

- (a) Determine ζ such that $f_i(\theta, \mathbf{Z}(Z_{k(i)} = \zeta)) = 0$. (To simplify the exposition, the algorithm will be written supposing that there is at most one root.) If it does not exist, skip this value of i .
- (b) Recompute $f_j(\theta, \mathbf{Z}(Z_{k(i)} = \zeta))$ for each $j \neq i$.
- (c) Compute $\frac{\partial f_i}{\partial(z_{k(i)}, \theta)}(\theta, \mathbf{Z}(Z_{k(i)} = \zeta))$.
- (d) Compute $\Delta^{(i)}\phi = \Delta^{(i)}\phi \left(\mathbf{I}_{f(\theta, \mathbf{Z}(Z_{k(i)} = \zeta)) > \mathbf{0}}, \theta, \mathbf{Z}(Z_{k(i)} = \zeta) \right)$ and increment the derivative:

$$\bar{\theta} += \Delta^{(i)}\phi \cdot \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \theta} \right] (\theta, \mathbf{Z}(Z_{k(i)} = \zeta)). \quad (3.16)$$

3. Compute the contribution of differentiable operations:

$$\bar{\theta} += \frac{\partial \phi}{\partial \theta}(\mathbf{I}_{f(\theta, \mathbf{Z}) > \mathbf{0}}, \theta, \mathbf{Z}). \quad (3.17)$$

Step **3** corresponds to addend D_0 in (3.11), and it is easily implemented with standard AAD, following backwards the smooth connections in the computational graph. In the same way one obtains $\partial f_i / \partial \theta$ and $\partial f_i / \partial z_{k(i)}$ in **2**(c).

Step **2** describes the computation of D_i , i.e. the contribution to the sensitivity caused by the i -th indicator function. The good news is that the computational graph for **2**(d) can be obtained from the original one with minor modifications. Indeed, for each solution $\zeta^{(i)}$ of equation $f_i(\theta, \mathbf{Z}(Z_{k(i)} = \zeta^{(i)})) = 0$, it is sufficient to:

- Substitute node $a_i = I_{f_i > 0}$ with a “node with two states” $a_i = \{0, 1\}$ no more depending on node f_i : the result obtained with $a_i = 0$ will be subtracted from the one obtained with $a_i = 1$.
- Substitute the random variable $Z_{k(i)}$ with $\zeta^{(i)}$.
- Substitute node f_i with a “weight” node (a rotated square in our pictures)

$$h_{\theta}(f_i, \mathbf{Z}) := \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \theta} \right] \quad (3.18)$$

connected to the result by a multiplicative link.

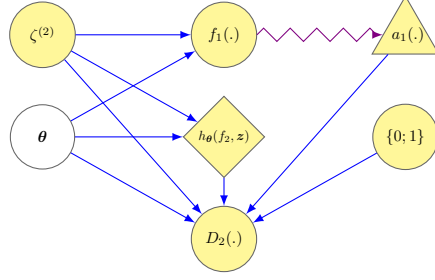


Figure 3.1: Computational graph for D_2 when $h = 2$ and $d = 1$. Yellow nodes must be recalculated, being directly or indirectly dependent on substituted nodes.

For instance, the computation of D_2 is described by the graph in Figure 3.1.

How much does this algorithm cost? Recall that thanks to algorithmic differentiation, the gradient of a function can be computed in a time which is a small multiple of the time spent computing the function itself, regardless of the length of the gradient. Therefore, the sum of the run times of steps **3** and **2(c)** is comparable to the original run time for the algorithm without derivatives, even when one wants to compute hundreds of sensitivities.

Step **2(b)** may be more critical, since it involves the re-evaluation of all f_j s for each $i = 1, \dots, h$; the same concerns may apply to **2(a)**, which requires multiple evaluations of f_i if the root is not analytically available, and finally to **2(d)**, which requires two evaluations of ϕ for each i . Hence we present a slight generalization of the algorithm, yielding in many cases a much smaller impact of these steps.

To this aim, in the concrete original implementation of the payoff, it is useful to identify which of the intermediate quantities that are computed do not depend on $Z_{k(i)}$. Let us denote with α_i the set of these variables (maybe inclusive of some components of θ and/or of $\mathbf{Z}$), and with f_j^i (resp. ϕ^i) the function which computes f_j (resp. ϕ) starting from α_i already computed:

$$f_j(\theta, \mathbf{Z}) := f_j^i(\alpha_i(\theta, \mathbf{Z}_{-k(i)}), Z_{k(i)}), \quad \phi(\mathbf{a}, \theta, \mathbf{Z}) := \phi^i(\mathbf{a}, \alpha_i(\theta, \mathbf{Z}_{-k(i)}), Z_{k(i)}). \quad (3.19)$$

Now one can observe that given functions α, f, g , the following holds:

$$\frac{\partial}{\partial \theta} \mathbb{E} [I_{f(\alpha(\mathbf{Z}_{-k}, \theta), Z_k) > 0} \cdot g(\mathbf{Z})] = \mathbb{E}_{\mathbf{Z}_{-k}} \left[\frac{\partial}{\partial \alpha} \mathbb{E}_{Z_k} [I_{f(\alpha(\mathbf{Z}_{-k}, \theta), Z_k) > 0} \cdot g(\mathbf{Z})] \cdot \frac{\partial \alpha}{\partial \theta} \right], \quad (3.20)$$

so that one can apply (3.6) to the inner expectation, getting derivatives with respect to α ; then the multiplication by $\partial \alpha / \partial \theta$ can be obtained implicitly by a suitably-initialized adjoint differentiation of the function $\alpha = \alpha(\theta)$.

By repeated application of this idea, algorithm 3.3 is modified as follows (Figure 3.2):

Algorithm 3.4. [Backward sweep with preprocessing] For each Monte Carlo scenario, after the forward sweep:

2. For $i = 1, \dots, h$:

- (a) Determine ζ such that $f_i^i(\alpha_i, \zeta) = 0$. If it does not exist, move to the next value of i .
- (b) Recompute $f_j^i(\alpha_i, \zeta)$ for each $j \neq i$.

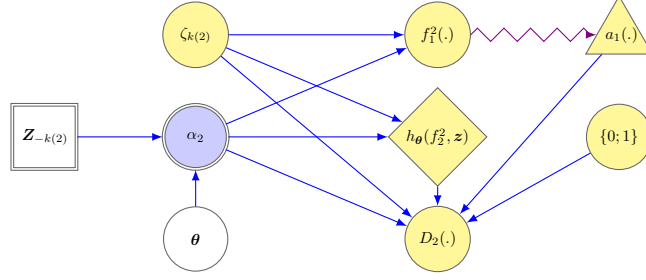


Figure 3.2: Computational graph for D_2 with preprocessing. In blue the portion of graph which does not need recalculation, being independent of the substituted nodes.

- (c) Compute $\nabla f_i^i(\alpha_i, \zeta)$ (with AAD).
- (d) Compute $\Delta^{(i)}\phi^i = \Delta^{(i)}\phi^i(\mathbf{I}_{\mathbf{f}^i(\alpha_i, \zeta) > \mathbf{0}}, \alpha_i, \zeta)$ and increment the derivative:

$$\bar{\alpha}_i += \Delta^{(i)}\phi^i \cdot \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i^i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i^i}{\partial \theta} \right] (\theta, \mathbf{Z}(Z_{k(i)} = \zeta)). \quad (3.21)$$

- 3. With $\bar{\alpha}_i$ initialized at the previous points and $\bar{\phi}$ initialized as 1, execute the backward sweep of the algorithm computing the payoff $\phi(\mathbf{I}_{\mathbf{f}(\theta, \mathbf{Z}) > \mathbf{0}}, \theta, \mathbf{Z})$ (considering the indicators and $\mathbf{Z}$ as constants).

In this new algorithm, the numerous recalculations of the f_j s and of ϕ , which could impact the performance, have been substituted by recalculations of the f_j^i s and ϕ^i s, hopefully much shorter.

Of course, to take advantage of this technique, one must identify α_i in some way. If an explicit representation of the computational graph is available in computer memory, then the minimal set of nodes β_i which must be recalculated when one “touches” $Z_{k(i)}$ can be determined by a simple tree traversal without computations, whose one-off cost is less than the cost of a single Monte Carlo scenario, and is therefore negligible when compared to the full pricing exercise. Then one can safely choose α_i as the complementary of the set β_i ; in fact some of these nodes might be unnecessary, but they do no harm, since thanks to algorithmic differentiation calculating a larger gradient causes no overhead.

On the other hand, if the implementation is not sufficiently introspective to allow for the building of such a graph, then the identification of α_i needs human intervention. Moreover, different algorithms to compute the same payoffs may benefit from the optimization described in this section to different degrees. In the next section we will show that when the risk factors are simulated via discretization of SDEs, the computations can be arranged in such a way that the efficiency gain is in fact relevant.

3.5 Monte Carlo

Suppose that the pricing model is described by a vector-valued diffusive process $\mathbf{X}^\theta$, whose dynamics can depend on parameters θ :

$$d\mathbf{X}^\theta(t) = \boldsymbol{\mu}(\theta, \mathbf{X}^\theta(t), t)dt + \boldsymbol{\Sigma}(\theta, \mathbf{X}^\theta(t), t)d\mathbf{W}_t, \quad \mathbf{X}^\theta(T_0) = \mathbf{X}_0(\theta), \quad (3.22)$$

where $\mathbf{W}$ denotes a standard N_W -dimensional Wiener process. Moreover, suppose that the payoff depends only on the value of $\mathbf{X}$ on a finite set of times $\mathbf{T} = \{T_1, \dots, T_M\}$, and maybe directly on $\boldsymbol{\theta}$:

$$P(\boldsymbol{\theta}) = \mathbb{E} [\phi(I_{f_1}(\boldsymbol{\theta}, \mathbf{X}^\theta(\mathbf{T})) > 0, \dots, I_{f_h}(\boldsymbol{\theta}, \mathbf{X}^\theta(\mathbf{T})) > 0, \boldsymbol{\theta}, \mathbf{X}^\theta(\mathbf{T}))]. \quad (3.23)$$

The most generic method to tackle the pricing problem (3.23) is to simulate $\mathbf{X}^\theta(\mathbf{T})$ discretizing the stochastic differential equation on a time grid $\{S_0, \dots, S_L\} = \mathbf{S} \supseteq \mathbf{T}$ with an Euler scheme:

$$\begin{aligned} \mathbf{X}^\theta(S_{i+1}) &= \mathbf{X}^\theta(S_i) + \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{X}^\theta(S_i), S_i) \Delta_i S + \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}^\theta(S_i), S_i) \sqrt{\Delta_i S} \cdot \mathbf{Z}^{(i)} \\ \Delta_i S &:= S_{i+1} - S_i, \quad \mathbf{Z}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d). \end{aligned} \quad (3.24)$$

This way, $\mathbf{X}^\theta(\mathbf{T})$ is expressed as a differentiable function of $\boldsymbol{\theta}$ and $\mathbf{Z} = (\mathbf{Z}^{(k)})_{k \geq 0}$ standard Gaussian vector, and hence (3.23) is in the form (3.8).

Algorithmically, this translates into a forward sweep of the following form:

Algorithm 3.5. [Forward sweep, SDE] For each Monte Carlo scenario:

1. Simulate random Gaussian vectors $\mathbf{Z}^{(0)}, \dots, \mathbf{Z}^{(L-1)}$.
2. Apply the Euler scheme (3.24), computing in particular the local drifts $\boldsymbol{\mu}^\theta(\mathbf{S} \setminus \{S_L\})$, the local diffusion coefficients $\boldsymbol{\Sigma}^\theta(\mathbf{S} \setminus \{S_L\})$, and the path $\mathbf{X}^\theta(\mathbf{S})$.
3. For $i = 1, \dots, h$: compute $f_i(\boldsymbol{\theta}, \mathbf{X}^\theta(\mathbf{T}))$.
4. Compute $\phi(I_{f(\boldsymbol{\theta}, \mathbf{X}^\theta(\mathbf{T})) > 0}, \boldsymbol{\theta}, \mathbf{X}^\theta(\mathbf{T}))$.

The corresponding computational graph is in Figure 3.3.

Note that each f_i will in general depend on the value of $\mathbf{X}$ on a subset of the fixing times $\mathbf{T}$: we call M_i the biggest element of such subset, and denote by $S_{t(i)}$ the time preceding M_i in the simulation grid $\mathbf{S}$, so that $M_i = S_{t(i)+1}$. Then to specify Algorithm 3.4 for this forward sweep, it suffices to:

- Call $\boldsymbol{\alpha}_i$ the collection of the following variables: $\mathbf{X}(S_0), \dots, \mathbf{X}(S_{t(i)}); \boldsymbol{\mu}^\theta(S_{t(i)}), \boldsymbol{\Sigma}^\theta(S_{t(i)}); \boldsymbol{\theta}; \mathbf{Z}_{-k(i)}^{(t(i))}; \mathbf{Z}^{(t(i)+1)}, \dots, \mathbf{Z}^{(L-1)}$; the value of f_j for each j such that $M_j < M_i$.
- Call $f_j^i(\boldsymbol{\alpha}_i, z)$ the function which computes f_j given $\boldsymbol{\alpha}_i$ and after imposing $Z_{k(i)}^{(t(i))} = z$. This function is trivial if $j \in J_i = \{j : M_j < M_i\}$, because the unaltered already-computed value of f_j is a component of $\boldsymbol{\alpha}_i$; while it includes the Euler steps from $S_{t(i)}$ to M_j if $M_j \geq M_i$ (i.e. $j \notin J_i$); in particular, it includes exactly one Euler step for $j = i$. Analogously, ϕ^i will include all the Euler steps from $S_{t(i)}$ to the contract's expiry.

We draw in Figure 3.4 the graph for D_i in the example $t(i) + 1 = L - 2$.

Remark 3.6 (Complexities). We know that the efficacy of the algorithm depends on the cost of the recalculations compared to the overall cost of pricing: let us estimate roughly these complexities.

The cost of pricing without sensitivities is evaluated as follows:

- The cost of a single Euler step will depend on the cost $C_{\boldsymbol{\mu}, \boldsymbol{\Sigma}}$ to compute $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and by the cost C_{Euler} of (3.24), dominated by the multiplication of $\boldsymbol{\Sigma}$ by $\mathbf{Z}$ which is of order $O(N_X N_W^2)$, where N_X is the dimension of $\mathbf{X}_t$. Anyway, the full simulation will have cost $L(C_{\boldsymbol{\mu}, \boldsymbol{\Sigma}} + C_{\text{Euler}})$.

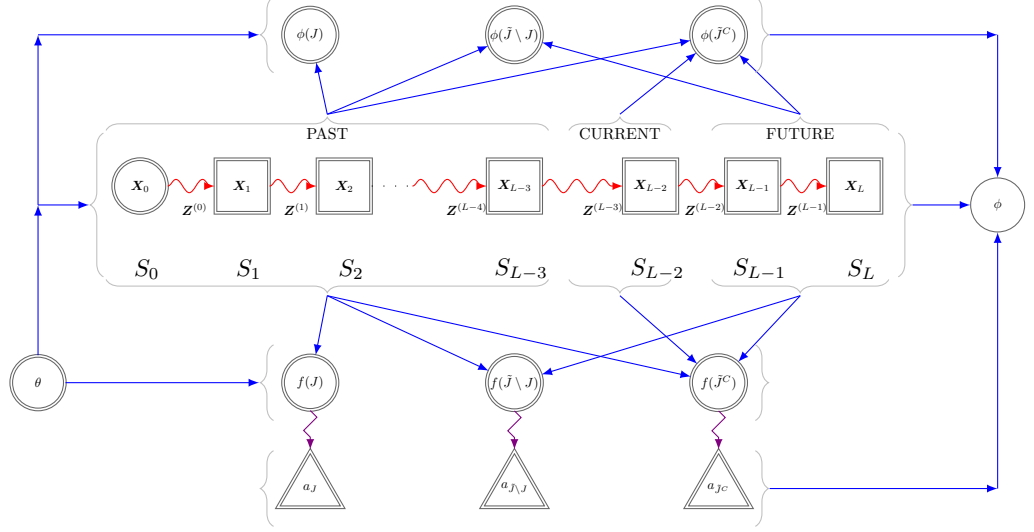


Figure 3.3: Graphical representation of the forward sweep for the calculation of the discounted payoff ϕ via Monte Carlo simulation of an SDE. To ease comparison with the following graphs for the computation of the derivative terms D_i , we split the functions f_j (i.e. the arguments of the indicator functions) into three sets.

For fixed i , the first set $f(J)$ contains those f_j s which depend only on fixing times before M_i (PAST). The second one $f(\tilde{J} \setminus J)$ contains those which can depend both on states before and after M_i , but not on the present state $X(M_i)$. The third set $f(\tilde{J}^C)$ collects all the remaining f_j s, i.e. those explicitly depending on $X(M_i)$.

With the same logic, we split the high portion of the computational graph, representing the algorithm to compute ϕ , into three portions $\phi(J)$, $\phi(\tilde{J} \setminus J)$ and $\phi(\tilde{J}^C)$.

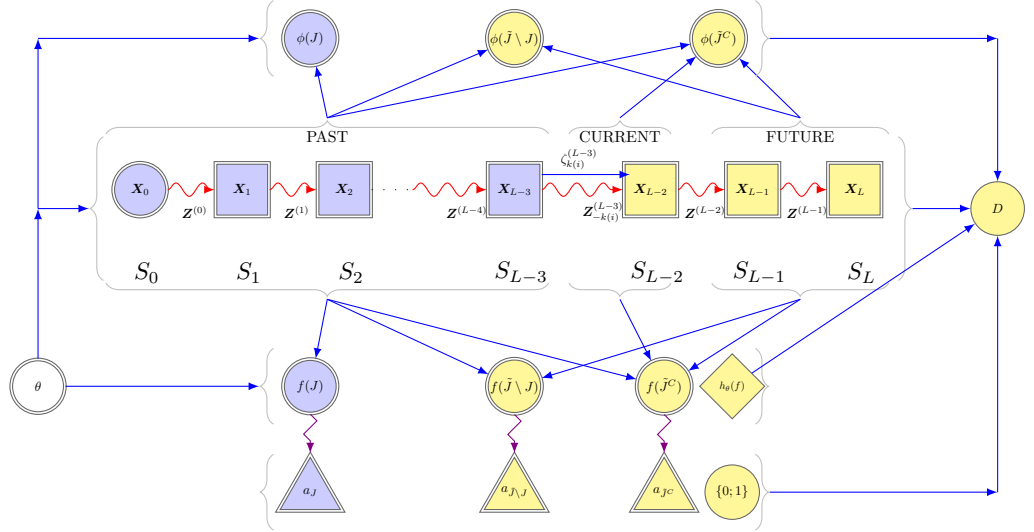


Figure 3.4: Graphical representation of the computation of the contribution D_i to the derivative; refer to the caption of Figure 3.3 for the organization of the graph. The computations which must not be repeated, represented by blue nodes, are those involving only past fixings (before M_i).

- The cost of the computation of the payoff given $\mathbf{X}$ will depend on the concrete ϕ ; we can decompose this complexity as $C_\phi + \sum_{i=1}^h C_{f_i}$, where C_ϕ is the complexity of the computation of ϕ given all f_j s, while C_{f_i} is the complexity of f_i .

The total cost is

$$L(C_{\mu, \Sigma} + C_{\text{Euler}}) + C_\phi + \sum_{i=1}^h C_{f_i}. \quad (3.25)$$

Only for the purpose of estimating the cost of recalculations, we suppose for simplicity that the discontinuities f_i have been ordered by maturity time (with no ties), and that for each i exactly one zero ζ_i is found. Then the i -th recalculation involves:

- $L - t(i)$ Euler steps from precalculated μ, Σ , each with cost C_{Euler} ;
- The recomputation of ϕ and f_j for $j \geq i$, which in the worst (unusual) case in which no intermediate result can be reused has a cost $C_\phi + \sum_{k=i}^h C_{f_k}$.

The total cost is

$$\sum_{i=1}^h (L - t(i)) C_{\text{Euler}} + h C_\phi + \sum_{i=1}^h i C_{f_i}. \quad (3.26)$$

Comparing (3.25) to (3.26), we see that we can have a problem only in one of the following cases:

1. There are several $t(i) \ll L$, so that the first addend in (3.26) becomes way larger than that in (3.25);
2. h is so large to be comparable with L , and either C_ϕ or some C_{f_i} for large i is significant, so that the second and/or third addend in (3.26), usually almost negligible, start playing a role.

To sum up, everything is fine if the M_i s are equal to the instrument's expiry $\max(\mathbf{T})$ or nearby. Instead, if there are digitalities involving only fixing times far from the instrument's maturity, the computation of α_i may consist of few operations, and proportionally the recalculations involved in step **2** of Algorithm 3.4 may be more heavyweight. However, it is possible to reformulate the simulation to reuse much more calculations. All one needs to note is that step **2** of Algorithm 3.5 can be substituted by:²

2bis.

- (a) Apply the Euler scheme (3.24) *on the reduced timegrid* $\mathbf{S}' = \mathbf{S} \setminus \mathbf{M} \cup \{S_L\}$, (hence some steps will go from S_t to S_{t+2} using the single random vector $\mathbf{Z}^{(t)}$), computing in particular the local drifts $\mu^\theta(\mathbf{S}' \setminus \{S_L\})$, the local diffusion coefficients $\Sigma^\theta(\mathbf{S}' \setminus \{S_L\})$, and the path $\mathbf{X}^\theta(\mathbf{S}')$.
- (b) For each $M_i = S_{t(i)+1} < S_L$, compute $\mathbf{X}(M_i)$ using a Brownian bridge:

$$\begin{aligned} \mathbf{X}(M_i) = & (1 - \lambda_{t(i)})\mathbf{X}(S_{t(i)}) + \lambda_{t(i)}\mathbf{X}(S_{t(i)+2}) + \\ & + \sqrt{(\Delta_{t(i)}S)(1 - \lambda_{t(i)})}\Sigma^\theta(S_{t(i)}) \cdot \mathbf{Z}^{t(i)+1}, \end{aligned} \quad (3.27)$$

where $\Delta_t S := S_{t+1} - S_t$ and $\lambda_t := (\Delta_t S + \Delta_{t+1} S)^{-1} \Delta_t S$.

²This discretization is indeed slightly different from the original one, but both converge to the exact SDE solution as the mesh ΔS tends to zero: more precisely, the distance in L^1 of each of the two discretizations from the true solution is $O(\Delta S^{1/2})$ (respectively Theorem 10.2.2 and Remark 10.2.3 in Kloeden and Platen, 1992), hence the distance between the two discretizations is also $O(\Delta S^{1/2})$. In any case, for maximum coherence, one could use the bridge-based scheme in both pricing and differentiation.

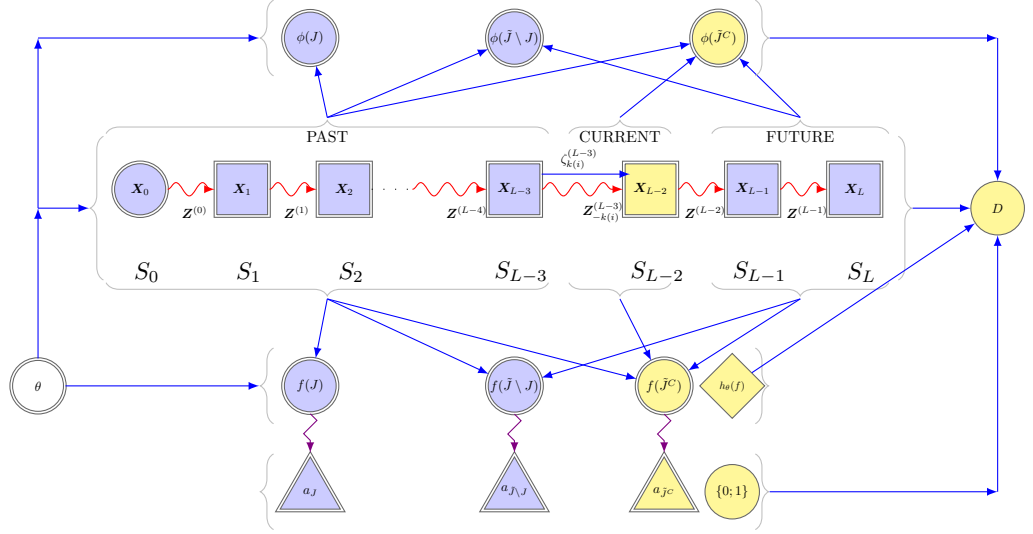


Figure 3.5: Graphical representation of the computation of the contribution D_i to the derivative in the Brownian bridge based variant. The number of nodes which are not recalculated (the blue nodes) is higher than in the standard formulation of Figure 3.4. In fact, one must recalculate only those nodes which depend directly or indirectly on fixing time M_i (CURRENT).

This way, Algorithm 3.4 can be applied with the following more advantageous definitions:

- Call α_i the collection of the following variables: $\mathbf{X}(S \setminus \{M_i\})$; $\mu^\theta(S_{t(i)})$, $\Sigma^\theta(S_{t(i)})$; θ ; $\mathbf{Z}_{-k(i)}^{(t(i)+1)}$; the value of f_j for all j such that f_j does not depend on $\mathbf{X}(M_i)$. We denote by $\tilde{J}_i$ the set of such j .
- Call $f_j^i(\alpha_i, z)$ the function which computes f_j given α_i and after imposing that $\mathbf{Z}_{-k(i)}^{(t(i)+1)} = z$. This function is trivial if $j \in \tilde{J}_i$, because the unaltered already-computed value of f_j is a component of α_i ; while if $j \in \tilde{J}_i^C$ (for instance if $j = i$), it involves simply:
 - (a) If $M_i < S_L$, a single Brownian bridge step (3.27);
 - (b) If $M_i = S_L$, the last Euler step (3.24).

This final algorithm is depicted in Figure 3.5.

Remark 3.7 (Complexities revisited). With the notation of Remark 3.6, we note that with the new algorithm the total complexity of recalculations will be

$$hC_{\text{bridge}} + hc_\phi + \sum_{i=1}^h n_i c_{f_i}, \quad (3.28)$$

where C_{bridge} is the cost of (3.27), n_i is the number of values of $j < i$ such that f_i depends on the exact fixing time M_j , and c_ϕ (resp. c_{f_i}) is the cost of updating ϕ (resp. f_i) with only one fixing perturbed. Now we note that in typical cases:

- $C_{\text{bridge}} \approx C_{\text{Euler}}$: the cost of a Brownian bridge step is of the same order of that of an Euler step.
- $c_\phi \ll C_\phi$ and $c_{f_i} \ll C_{f_i}$: updating ϕ or some f_i for a pointwise path perturbation is much less costly than computation from scratch.
- $n_i \ll i$: most discontinuities depend only on some fixings.

In view of these facts, the improvement with respect to (3.26) is evident: when compared to the cost of pricing (3.25), the new cost (3.28) is almost always small, and for the hardest payoffs anyway not large.

3.6 Numerical examples

In this section, we will apply the method described in the rest of the chapter to financial products which are combinations of indicator functions $H(x) := I_{x>0}$. For the sake of brevity, we will refer to the method itself with the acronym DAAD (“Distributional” of “Dirac-delta” AAD).

To avoid unneeded complications, all the examples will involve underlyings A with geometric Brownian motion dynamics in the risk-neutral measure:

$$\frac{dA_t}{A_t} = \mu(t)dt + \sigma(t)dW_t, \quad (3.29)$$

where μ and σ are deterministic scalar functions, and W_t is a standard Wiener process with respect to the filtration $(\mathcal{F}_t)_{t \geq T_0}$ representing market information up to time t . For such a simple choice, the law of A_T would be known in closed form, but all Monte Carlo estimators will not make use of this peculiarity of the model, to make the analysis hopefully indicative of what one would get for less trivial dynamics.

We will suppose that the law of A is independent of risk-free interest rates $r(t)$, so that the fair price at time t of a cashflow π_T paid at time $T > t$ and depending only on A will be (see for instance Duffie, 2001; Björk, 2009):

$$P_t = E[D(t, T)\pi_T \mid \mathcal{F}_t] = P(t, T)E[\pi_T \mid \mathcal{F}_t], \quad (3.30)$$

where we introduced the notations

$$D(t, T) = \exp\left(-\int_t^T r(s)ds\right), \quad P(t, T) = E[D(t, T) \mid \mathcal{F}_t]. \quad (3.31)$$

Finally, all run times are based on a C++ implementation run on a personal computer with an Intel(R) Xeon(R) CPU E5-2650 @ 2.00 GHz. Anyway, while the actual number of seconds may differ according to hardware and software details, we have found the ratio between different methods to be fairly stable across machines and changes in the implementation.

3.6.1 Digital call

Let us start with the case of a digital (or binary) call with strike K on the value A_T of the underlying at a single time T : in formulas, the discounted payout is

$$\phi_T = D(t, T)H(A_T - K), \quad (3.32)$$

where we recall that $H(x) = I_{x>0}$ denotes the Heaviside function. Given the dynamics (3.29), price B_t and sensitivities of the binary call are known in closed form (see e.g.

Björk, 2009). We will mainly concentrate on the option's Delta sensitivity, i.e. the sensitivity to the initial level A_t , but the analysis can be easily repeated for the other sensitivities (we recall that the method is specifically designed to compute large numbers of them).

We want to test Monte Carlo methods in this lucky case in which analytical values are available. To this purpose, we simulate N paths $(\omega_i)_{i=1,\dots,N}$ of asset A and firstly estimate the expected value defining the price as a sample mean of the realizations $\phi_T(\omega_i)$ of the payoff ϕ_T , i.e.

$$B_t^{(N)} = \frac{1}{N} \sum_{i=1}^N \phi_T(\omega_i). \quad (3.33)$$

Analogously, we estimate the Delta using the results of Section 3.5 with $\phi \equiv 1$ and $f_1 := A_T - K$:

$$\left(\frac{\partial B_t}{\partial A_t} \right)^{(N)} = \frac{1}{N} \sum_{i=1}^N D(t, T) \frac{e^{-\zeta_{t(1)}^2/2}}{\sqrt{2\pi}} \left[\frac{\partial A_T}{\partial A_t} \left| \frac{\partial A_T}{\partial z} \right|^{-1} \right] (\tilde{\omega}_i), \quad (3.34)$$

where $\tilde{\omega}_i$ is the perturbed trajectory in which the Brownian increment $Z_{t(1)}$ has been substituted with the root $\zeta_{t(1)}$ of the appropriate equation (see Figure 3.6). With $p := t(1)$ (index of the time in the mesh preceding the maturity $M_1 = T$), we note that in the Euler-discretized Black-Scholes model

$$\frac{\partial A_T}{\partial A_t} = \frac{\partial A_T}{\partial A_{S_p}} \frac{\partial A_{S_p}}{\partial A_t} = \frac{A_T}{A_{S_p}} \frac{A_{S_p}}{A_t} = \frac{A_T}{A_t}, \quad \frac{\partial A_T}{\partial z} = \sqrt{T - S_p} A_{S_p} \sigma(S_p); \quad (3.35)$$

substituting these expressions into (3.34) and noting that $A_T(\tilde{\omega}_i) = K$ we get:

$$\left(\frac{\partial B_t}{\partial A_t} \right)^{(N)} = \frac{1}{N} \sum_{i=1}^N D(t, T) \frac{K}{A_t} \frac{e^{-\zeta_p^2/2}}{\sqrt{2\pi}} \left| \sqrt{T - S_p} A_{S_p} \sigma(S_p) \right|^{-1} (\omega_i). \quad (3.36)$$

We list in Table 3.1 the results obtained for the price, Delta and Vega³ sensitivity of *amount* = 10000 binary call options with $A_0 = 100$, $K = 100$, $T = 5$, $\mu(t) = \mu = 1\%$, $\sigma(t) = \sigma = 10\%$, for different values of N and of the discretization step $\Delta_i S = S_{i+1} - S_i$. Here and in the following tables, the uncertainties (Unc) are computed as 2.33 times the standard deviation of the Monte Carlo average, so that they represent the half-width of a 98%-confidence interval according to the asymptotic distribution of such average, which is normal by the central limit theorem.

Comparison with approximate replications by call spreads

This paragraph performs an in-depth comparison of our method to a more straightforward one which is popular among practitioners. Once more, the procedure is described and detailed in the simplest case of a digital call, but it will serve as benchmark without further explanations also in the subsequent examples.

For the purpose of sensitivity calculation, a common trick is to approximate the binary call with strike K using a special combination of vanilla options, precisely a long call with strike $K_- := K - \varepsilon/2$ and a short call with strike $K_+ := K + \varepsilon/2$ where $\varepsilon > 0$: such combination is named *call spread* or *vertical spread*. The structure

³Vega is the common name of the sensitivity with respect to an uniform bump of $\sigma(t)$.

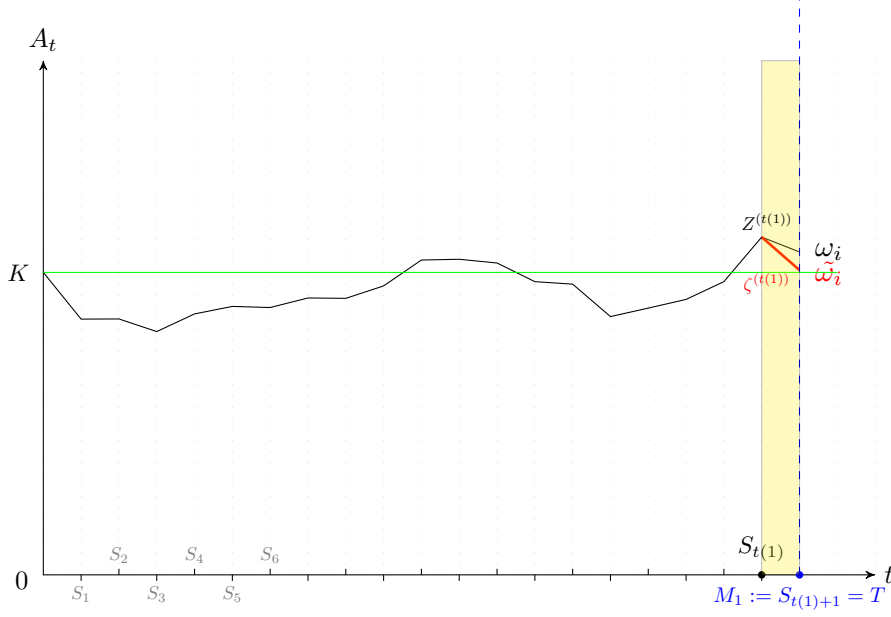


Figure 3.6: Path simulation in digital call.

of the resulting payoff is analogous to that of the digital, provided that one uses ε^{-1} call spreads to replicate the unitary height of the Heaviside step. The approximation is increasingly good as $\varepsilon \rightarrow 0$: indeed, while ε decreases, the slope of the non-flat section increases, approaching, in the limit, the vertical direction which is typical of the binary's payoff.

The call spread is a Lipschitz payoff, hence (Glasserman, 2004) notwithstanding the non-differentiability in $K \pm \varepsilon/2$, standard pathwise algorithmic differentiation (AAD) is applicable. Note that the price $CS_t(A_t, K, \varepsilon)$ at time t of the call spread is just a combination of call prices, so that we can compute analytically the biases introduced by the regularization in the value and Delta:

$$\text{Bias} \{CS_t(A_t, K, \varepsilon)\} := CS_t(A_t, K, \varepsilon) - B_t, \quad (3.37)$$

$$\text{Bias} \{\delta CS_t(A_t, K, \varepsilon)\} := \delta CS_t(A_t, K, \varepsilon) - \frac{\partial B_t}{\partial A_t}. \quad (3.38)$$

By construction, the approximation of the digital payoff will be better if the regularization parameter ε is smaller: indeed

$$\lim_{\varepsilon \rightarrow 0} \text{Bias} \{CS_t(A_t, K, \varepsilon)\} = \lim_{\varepsilon \rightarrow 0} \text{Bias} \{\delta CS_t(A_t, K, \varepsilon)\} = 0. \quad (3.39)$$

This is not equally true for the Delta's pathwise estimator $\delta CS_t^{(N)}(A_t, K, \varepsilon)$, because the variance of the Monte Carlo average $\delta CS_t^{(N)}(A_t, K, \varepsilon)$ is higher if ε is smaller:

$$\text{Var} [\delta CS_t^{(N)}(A_t, K, \varepsilon)] = N^{-1} O(\varepsilon^{-1}).$$

(Appendix A.3 derives the actual expression, displayed in equation (A.40).)

To sum up, there is a tradeoff between the estimator's bias, diminishing if ε is smaller, and its Monte Carlo variance, which for a fixed number of simulations has

Table 3.1: Price, Delta and Vega sensitivity of a digital call. The analytical value is compared to several DAAD estimates, differing by number N of simulated paths and by length $\Delta_i S$ of the Euler steps. The uncertainties represent a 98%-confidence interval.

Quantity	Value	step ($\Delta_i S$)	MC scen (N)	MC estimate	Unc
Price	5546.11	400^{-1}	10^4	5490.64	118.45
		100^{-1}	10^4	5425.80	118.48
		400^{-1}	10^5	5555.98	37.42
		100^{-1}	10^5	5529.74	37.43
		400^{-1}	10^6	5551.09	11.83
		100^{-1}	10^6	5556.30	11.83
Delta	180.59	400^{-1}	10^4	179.93	23.46
		100^{-1}	10^4	176.80	16.11
		400^{-1}	10^5	180.99	7.41
		100^{-1}	10^5	180.19	5.13
		400^{-1}	10^6	181.09	2.34
		100^{-1}	10^6	179.97	1.62
Vega	-13544.23	400^{-1}	10^4	-13495.01	1759.68
		100^{-1}	10^4	-13260.32	1208.34
		400^{-1}	10^5	-13574.91	555.88
		100^{-1}	10^5	-13514.50	384.98
		400^{-1}	10^6	-13582.01	175.61
		100^{-1}	10^6	-13497.40	121.68

the opposite trend. To quantify this tradeoff, we consider the mean square error I of the estimate obtained by regularization with parameter ε , which is equal to the quadratic sum of bias and statistical uncertainty:

$$I(\varepsilon) = \sqrt{\text{Var}[\delta C S_t^{(N)}(A_t, K, \varepsilon)] + \text{Bias}^2\{\delta C S_t(A_t, K, \varepsilon)\}}. \quad (3.40)$$

We will choose the optimal ε minimizing this error metric and call it $\bar{\varepsilon}$ (note that this would be highly nontrivial for most payoffs).

We put in Table 3.2 a comparison between the estimates of the Delta sensitivity obtained with (3.36) [DAAD] and those based on the call spread approximation [Smoothing] with the optimal $\bar{\varepsilon}$, for a binary call with $A_0 = 100$, $K = 100$, $\mu(t) = \mu = 1\%$, $\sigma(t) = \sigma = 10\%$, *amount* = 10000, Monte Carlo parameters $\Delta_i S = 0.01$ and $N = 10^6$, for different expiries T . In this example, the DAAD method looks especially good for short-lived options, and has statistical uncertainties never worse than those of the $\bar{\varepsilon}$ -smoothed estimator; however, to the latter ones we should sum the bias.

The same analysis is repeated for the Vega sensitivity in Table 3.3 (see Appendix A.3 for detailed computations). Here the advantage of DAAD is even more relevant, unless the option is very long-lived. Note also that usually one would not use a different ε for each sensitivity, to avoid running the pricing many times: therefore, in practice ε would be a compromise between $\bar{\varepsilon}_\delta$ and $\bar{\varepsilon}_\nu$. However, the two numbers are not so different at least in this example.

Table 3.2: Delta sensitivity of a digital call for different maturities T (expressed in years). The estimated value and its 98%-confidence Monte Carlo uncertainty are computed both by DAAD and by a smoothing approach with optimal width $\bar{\varepsilon}_\delta$. The analytical reference value and theoretical bias of the smoothing method are also reported.

T	$\partial B_t / \partial A_t$	Smoothing					DAAD		
		$\bar{\varepsilon}_\delta$	Val	Unc	Bias	runtime	Val	Unc	runtime
0.03	2410.37	0.34	2413.24	18.80	-4.22	22	2411.15	6.39	22
0.05	1796.67	0.46	1790.86	13.93	-3.14	26	1796.02	4.55	26
0.1	1270.58	0.64	1264.92	9.93	-2.22	38	1272.90	4.23	38
0.5	565.64	1.45	561.87	4.40	-0.99	108	565.07	2.16	108
1	400.01	2.04	400.21	3.13	-0.7	182	400.94	2.30	182
5	180.59	4.55	180.016	1.42	-0.32	844	179.97	1.62	846
10	123.29	6.45	123.05	0.98	-0.22	1636	122.18	1.33	1638

Table 3.3: Vega sensitivity of a digital call for different maturities T (expressed in years). The estimated value and its 98%-confidence Monte Carlo uncertainty are computed both by DAAD and by a smoothing approach with optimal width $\bar{\varepsilon}_\nu$. The analytical reference value and theoretical bias of the smoothing method are also reported.

T	$\partial B_t / \partial \sigma$	Smoothing					DAAD		
		$\bar{\varepsilon}_\nu$	Val	Unc	Bias	Time	Val	Unc	Time
0.03	-990.56	0.14	-997.25	17.51	0.3	22	-990.89	2.63	22
0.05	-1329.04	0.25	-1317.88	20.08	0.70	26	-1328.56	3.37	26
0.1	-1879.76	0.46	-1859.6	23.79	1.70	38	-1883.2	6.26	38
0.5	-4230.67	1.35	-4205.43	38.85	6.54	108	-4226.38	16.17	108
1	-6000.1	1.98	-5994.39	51.26	9.89	182	-6014.09	34.57	182
5	-13544.23	4.53	-13506.2	108.20	23.57	844	-13497.4	121.68	846
10	-18493.92	6.41	-18456.5	147.89	32.53	1636	-18327.6	200.48	1638

3.6.2 Double digital call

Let us consider as a second example an option consisting of a single payment of 1 at time T recognized only if the fixing of the underlying A at two fixing times F_1 and F_2 is higher than K_1 and K_2 respectively, with $F_1 < F_2 \leq T$. Supposing for the sake of simplicity that $F_2 = T$, the discounted payoff is

$$\phi_T = D(t, T) H(A_{F_1} - K_1) H(A_T - K_2). \quad (3.41)$$

Its price $BB_t(K_1, K_2)$ at time t cannot be solved analytically. Since thanks to the tower rule we can write

$$BB_t(K_1, K_2) = \mathbb{E}[D(t, F_1) H(A_{F_1} - K_1) B_{F_1}(K_2) \mid \mathcal{F}_t] =: \mathbb{E}[\tilde{\phi}_T \mid \mathcal{F}_t], \quad (3.42)$$

which is the price of a more tractable single-discontinuity payoff, we will take as benchmark for our analysis the following sample-based estimates: for the price

$$BB_t^{(N)}(K_1, K_2) := \frac{1}{N} \sum_{i=1}^N D(t, F_1, \omega_i) H(A_{F_1}(\omega_i) - K_1) B_{F_1}(K_2, \omega_i), \quad (3.43)$$

and for the Delta sensitivity the DAAD estimator applied to the partially-integrated discounted payoff $\tilde{\phi}_T$:

$$\begin{aligned} \delta BB_t^{(N)}(K_1, K_2) &:= \frac{1}{N} \sum_{i=1}^N D(t, F_1, \omega_i) H(A_{F_1}(\omega_i) - K_1) \frac{\partial B_{F_1}(K_2, \omega_i)}{\partial A_{F_1}} \frac{\partial A_{F_1}}{\partial A_t} \\ &\quad + \frac{1}{N} \sum_{i=1}^N D(t, F_1, \tilde{\omega}_i) \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\frac{\partial A_{F_1}}{\partial A_t} \left| \frac{\partial A_{F_1}}{\partial z} \right|^{-1} B_{F_1}(K_2) \right] (\tilde{\omega}_i). \end{aligned} \quad (3.44)$$

Let us write down the estimator obtained if we apply the DAAD method to the original payoff ϕ_T instead. The second indicator function depends on the value of A at maturity, hence its contribution to the derivative is simply (defining $p_i := t(i)$):

$$\begin{aligned} \delta^{(2)} BB_t^{(N)}(K_1, K_2) &:= \\ &= \frac{1}{N} \sum_{i=1}^N D(t, T) H(A_{F_1}(\omega_i) - K_1) \frac{e^{-\zeta_{p_2}^2/2}}{\sqrt{2\pi}} \left[\frac{\partial A_T}{\partial A_{S_{p_2}}} \frac{\partial A_{S_{p_2}}}{\partial A_t} \left| \frac{\partial A_{F_2}}{\partial z} \right|^{-1} \right] (\tilde{\omega}_i) \\ &= \frac{1}{N} \sum_{i=1}^N D(t, T) H(A_{F_1}(\omega_i) - K_1) \frac{e^{-\zeta_{p_2}^2/2}}{\sqrt{2\pi}} \left[\frac{K_2}{A_t} \left| \sqrt{T - S_{p_2}} A_{S_{p_2}} \sigma(S_{p_2}) \right|^{-1} \right] (\omega_i), \end{aligned} \quad (3.45)$$

where $\tilde{\omega}_i$ is the perturbed path obtained substituting the Brownian increment $Z_{t(2)}$ with the root $\zeta_{t(2)}$ of the appropriate equation (see Figure 3.7).

For the intermediate indicator function we apply the technique based on Brownian bridge, writing

$$A_{F_1} = [\bar{\lambda}_{p_1} + \Sigma(p_1) Z_{p_1+1}] A_{S_{p_1}} + \lambda_{p_1} A_{S_{p_1+2}} \quad (3.46)$$

with $\bar{\lambda}_{p_1} := (1 - \lambda_{p_1})$ and $\Sigma(p_1) := \sqrt{(F_1 - S_{p_1}) \bar{\lambda}_{p_1} \sigma(S_{p_1})}$.

The corresponding contribution to the derivative is therefore

$$\begin{aligned} \delta^{(1)} BB_t^{(N)}(K_1, K_2) &:= \\ &= \frac{1}{N} \sum_{i=1}^N D(t, T) H(A_{F_2}(\omega_i) - K_2) \frac{e^{-\zeta_{p_1+1}^2/2}}{\sqrt{2\pi}} \times \\ &\quad \times \left[\left(\frac{\partial A_{F_1}}{\partial A_{S_{p_1+2}}} \frac{\partial A_{S_{p_1+2}}}{\partial A_t} + \frac{\partial A_{F_1}}{\partial A_{S_{p_1}}} \frac{\partial A_{S_{p_1}}}{\partial A_t} \right) \left| \frac{\partial A_{F_1}}{\partial z} \right|^{-1} \right] (\tilde{\omega}_i) \\ &= \frac{1}{N} \sum_{i=1}^N D(t, T) H(A_{F_2}(\omega_i) - K_2) \frac{e^{-\zeta_{p_1+1}^2/2}}{\sqrt{2\pi}} \times \\ &\quad \times \frac{1}{A_t} \left[(\lambda_{p_1} A_{S_{p_1+2}} + [\bar{\lambda}_{p_1} + \Sigma(p_1) \zeta^{p_1}] A_{S_{p_1}}) |\Sigma(p_1) A_{S_{p_1}}|^{-1} \right] (\tilde{\omega}_i) \\ &= \frac{1}{N} \sum_{i=1}^N D(t, T) H(A_{F_2}(\omega_i) - K_2) \frac{e^{-\zeta_{p_1+1}^2/2}}{\sqrt{2\pi}} \frac{K_1}{A_t} |\Sigma(p_1) A_{S_{p_1}}|^{-1} (\omega_i), \end{aligned} \quad (3.47)$$

where $\tilde{\omega}_i$ is the perturbed path where the Brownian increment $z_{t(1)+1}$ is substituted by the appropriate root $\zeta_{t(1)+1}$ (see again Figure 3.7).

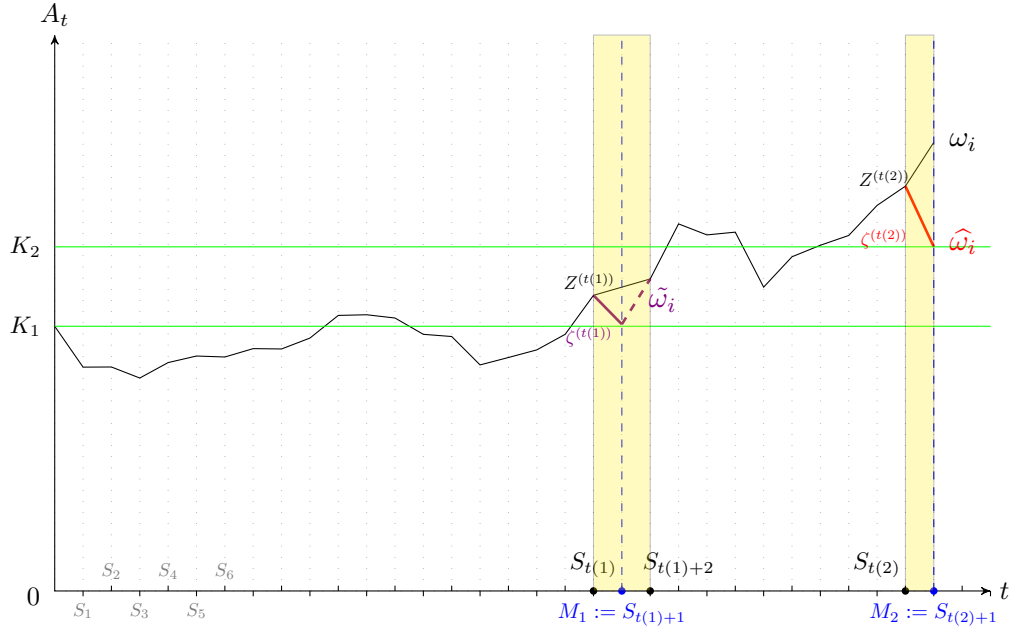


Figure 3.7: Path simulation with Brownian bridge in double digital call.

Turning to the numbers, we take an *amount* = 10000 of 2-years-long double digitals with strikes $K_1 = K_2 = 100$ and various fixing times F_1 , and compare the benchmarks (3.43) and (3.44) with the estimates given by DAAD and those obtained smoothing the two indicator functions. For the latter we do not have an exact way to compute the optimal smoothing parameters, so we put $\bar{\varepsilon}_i$ equal to the optimal smoothing parameter of a digital with strike K_i expiring at time F_i , which should be a reasonable guess. The underlying's model is defined by $A_0 = 100$, $\mu(t) = \mu = 1\%$, $\sigma(t) = \sigma = 10\%$. The simulation time-step is $\Delta_i S = 400^{-1}$ and the number of scenarios is $N = 10^6$. The prices are in Table 3.4, the Delta sensitivity in Table 3.5, and the Vega sensitivity in Table 3.6. The reader can appreciate from those numbers that also for the double digital, the DAAD method is competitive.

Table 3.4: Monte Carlo prices of double digital calls paying $H(A_{F_1} - K_1)H(A_{F_2} - K_2)$ at time F_2 for different values of F_1 , F_2 (expressed in years). A benchmark value is compared to the results obtained in runs where sensitivities were also computed, respectively by smoothing of the payoff or by DAAD. Computational times in seconds of both methods are reported.

F_1	F_2	Benchmark		Smoothing					DAAD		
		Val	Unc	$\bar{\varepsilon}_1$	$\bar{\varepsilon}_2$	Val	Unc	Time	Val	Unc	Time
1	2	4038.54	9.32	2.04	2.89	4034.46	11.18	98	4039.71	11.52	98
0.1	2	3060.96	7.09	0.64	2.89	3061.79	10.45	98	3061.81	10.81	98
0.01	2	2791.17	6.48	0.19	2.89	2794.73	10.22	98	2794.33	10.51	98

If, instead of choosing the two smoothing parameters for the two indicator functions independently, we had taken a single parameter for both fixings, the results for price and Delta would have been those in Tables 3.7 and 3.8. The boldface lines are those

Table 3.5: Delta sensitivity of double digital calls paying $H(A_{F_1} - K_1)H(A_{F_2} - K_2)$ at time F_2 for different values of F_1, F_2 (expressed in years). A benchmark value is compared to the results obtained by smoothing of the payoff or by DAAD. Computational times in seconds of both methods are reported.

F_1	F_2	Benchmark		Smoothing					DAAD		
		Val	Unc	$\bar{\varepsilon}_1$	$\bar{\varepsilon}_2$	Val	Unc	Time	Val	Unc	Time
1	2	351.33	1.84	2.04	2.89	351.27	2.79	346	352.76	3.66	352
0.1	2	814.82	3.63	0.64	2.89	817.31	7.50	358	814.50	5.59	360
0.01	2	2268.93	5.13	0.19	2.89	2267.11	24.09	345	2271.64	9.18	352

Table 3.6: Vega sensitivity of double digital calls paying $H(A_{F_1} - K_1)H(A_{F_2} - K_2)$ at time F_2 for different values of F_1, F_2 (expressed in years). A benchmark value is compared to the results obtained by smoothing of the payoff or by DAAD. Computational times in seconds of both methods are reported.

F_1	F_2	Benchmark		Smoothing					DAAD		
		Val	Unc	$\bar{\varepsilon}_1$	$\bar{\varepsilon}_2$	Val	Unc	Time	Val	Unc	Time
1	2	-7427.74	29.72	2.04	2.89	-7283.22	60.24	346	-7433.7	77.16	352
0.1	2	-5271.51	10.81	0.64	2.89	-5249.07	51.91	358	-5273.77	62.58	360
0.01	2	-4581.89	9.90	0.19	2.89	-4589.54	51.03	345	-4595.98	62.22	352

where the smoothed estimator's confidence interval fails to contain the benchmark. This exercise makes apparent that the choice of good smoothing parameters needs care, and may be problematic when the discontinuities are many and/or the payoff is more complex.

Table 3.7: Monte Carlo price of $H(A_{F_1} - K_1)H(A_{F_2} - K_2)$ paid at time F_2 for different values of F_1, F_2 , where both discontinuities have been smoothed using the same parameter $\varepsilon_1 = \varepsilon_2$.

F_1	F_2	Benchmark		Smoothing			
		Val	Unc	ε_1	ε_2	Val	Unc
0.1	2	3060.96	7.09	2.89	2.89	3049.79	9.79
				0.64	0.64	3061.33	10.58
0.01	2	2791.17	6.48	2.89	2.89	2763.83	7.79
				0.19	0.19	2794.66	10.33

3.6.3 Barrier options

Next we consider the case of barrier options. Their payoff depends on the underlying touching or not a level or *barrier*. We will focus on options of type “down and out”, i.e. which expire with zero payout when the price of the underlying A goes below a (possibly time-dependent) barrier; if this never happens till maturity T , a cashflow $\phi(A)$ is paid at a time T_P . Denoting by $(T_i)_{i=1,\dots,n}$ the discrete timegrid where the barrier breaches are monitored, the discounted payoff is therefore

$$\phi_{T_P} = D(t, T_P) \tilde{\phi}(A_{T_n}) \prod_{i=1}^n H(A(T_i) - K_i). \quad (3.48)$$

Table 3.8: Pathwise Delta of $H(A_{F_1} - K_1)H(A_{F_2} - K_2)$ paid at time F_2 for different values of F_1, F_2 , where both discontinuities have been smoothed using the same parameter $\varepsilon_1 = \varepsilon_2$.

F_1	F_2	Benchmark		Smoothing			
		Val	Unc	ε_1	ε_2	Val	Unc
0.1	2	814.82	3.63	2.89	2.89	792.38	3.45
				0.64	0.64	819.56	8.17
0.01	2	2268.93	5.13	2.89	2.89	1714.65	4.15
				0.19	0.19	2267.54	25.03

To keep it simple, we will evaluate the Delta sensitivity in the case in which $\tilde{\phi}(A_{T_n}) = 1$. The underlying dynamics's parameters are set to $A_0 = 100$, $\mu(t) = \mu = 1\%$, $\sigma(t) = \sigma = 10\%$. The simulation time-step is $\Delta_i S = 400^{-1}$, and the Monte Carlo scenarios are $N = 10^6$.

Figures 3.8 and 3.9 report the results for $n = 20$ barrier monitoring times, while Figures 3.10 and 3.11 refer to $n = 40$. In all these figures, in the table on the left, the first line represents a benchmark obtained by finite difference with a much higher number 10^8 of scenarios; the second row is the DAAD estimate; and the subsequent lines are obtained with pathwise differentiation of a smoothed payoff, for different values of the smoothing parameter ε . Computational times are also reported. The right half of each figure plots the same results (values and statistical uncertainties) in relative units with respect to the benchmark (the green line); the x-axis runs through the line numbers in the table on the left, in particular the blue point is obtained by DAAD while the red ones are obtained by smoothing. Finally, the captions contain the details of the four experiments and the only-price run time.

We can make the following remarks:

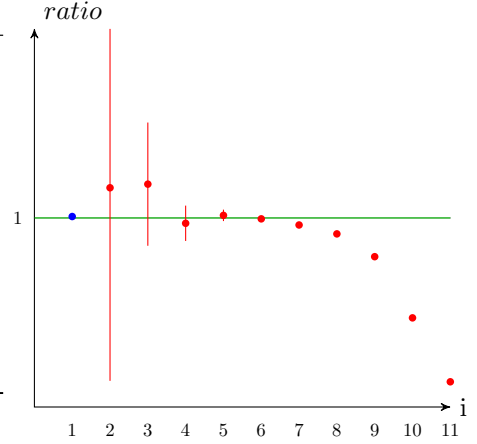
1. The DAAD estimate is always in line with the reference value;
2. The choice of the regularization parameter is crucial to obtain, with the “Smoothing” method, both correct results and acceptable uncertainties. The optimal value is difficult to predict a-priori and depends on the payoff and on the sensitivity of interest;
3. The uncertainty of the DAAD estimate is always of the same dimension of the uncertainty obtained with the empirically-optimal ε ;
4. The execution times for DAAD are comparable to those of the evaluation via smoothing. The actual numbers of course depend on how many indicator functions are involved. In the worst of the examined cases (40 monitoring times), the time for DAAD is approximately twice that for smoothing.

3.6.4 Digital on time average

We repeat the analysis of Section 3.6.3 for an instrument paying at time T_P an amount equal to 1 only if the average of the values of the underlying A at times $T_i \leq T_P$ is above a certain strike K . The discounted payoff can be written this way:

$$\phi = D(t, T_P) H\left(\frac{1}{n} \sum_{i=1}^n A(T_i) - K\right). \quad (3.49)$$

i	Method	ϵ	δ Avg	δ Unc	Time
0	Ref.	0	608.20	1.28	34177
1	DAAD	0	612.74	5.79	672
2	Smooth.	0.0001	705.25	621.10	478
3	Smooth.	0.001	716.92	198.24	478
4	Smooth.	0.01	590.99	56.87	478
5	Smooth.	0.1	616.50	18.39	478
6	Smooth.	1	605.25	5.81	478
7	Smooth.	2	585.55	4.11	478
8	Smooth.	3	556.92	3.34	478
9	Smooth.	5	483.56	2.53	478
10	Smooth.	10	286.75	1.50	478
11	Smooth.	20	81.10	0.57	478



i	Method	ϵ	ν Avg	ν Unc	Time
0	Ref.	0	-2232.47	7.81	34191
1	DAAD	0	-2218.84	32.38	672
2	Smooth.	0.0001	-2272.14	2329.35	478
3	Smooth.	0.001	-2707.52	1141.85	478
4	Smooth.	0.01	-2220.35	321.35	478
5	Smooth.	0.1	-2252.6	101.82	478
6	Smooth.	1	-1861.14	31.93	478
7	Smooth.	2	-920.15	24.29	478
8	Smooth.	3	366.92	24.03	478
9	Smooth.	5	3087.87	30.63	478
10	Smooth.	10	6235.69	42.07	478
11	Smooth.	20	4011.26	33.29	478

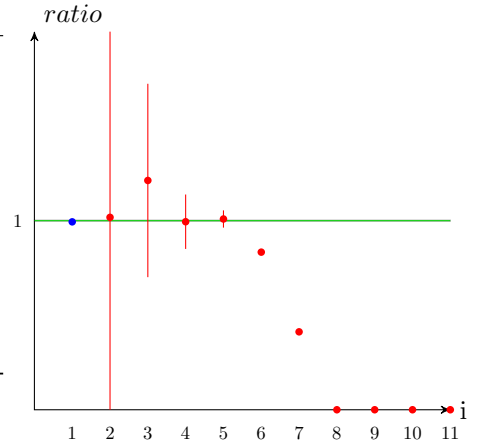
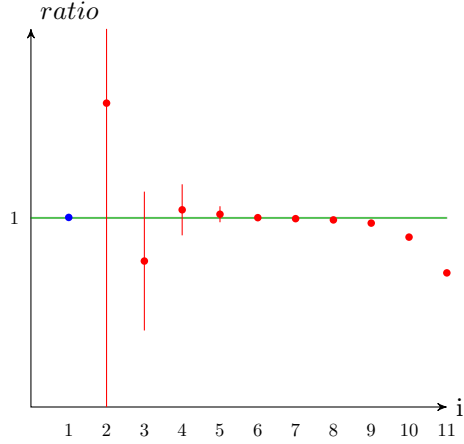


Figure 3.8: Delta and Vega sensitivity computed with DAAD and with ϵ -wide smoothing for the payoff $\phi = \prod_{i=1}^n H(A(T_i) - K_i)$ with $n = 20$, $T_i = i * 0.05$, strike $K_i = K = 100$, payment time $T_P = 2$. The green line (experiment $i = 0$) is a benchmark obtained by finite difference with a much higher number of scenarios. The only-price run time was 171 seconds.

i	Method	ϵ	δ Avg	δ Unc	Time
0	Ref.	0	313.27	0.92	34811
1	DAAD	0	314.13	4.13	909
2	Smooth.	0.0001	503.51	524.67	671
3	Smooth.	0.001	241.83	115.02	671
4	Smooth.	0.01	326.74	42.29	671
5	Smooth.	0.1	319.40	13.25	671
6	Smooth.	1	313.72	4.24	671
7	Smooth.	2	312.07	3.08	671
8	Smooth.	3	309.99	2.597	671
9	Smooth.	5	304.71	2.13	671
10	Smooth.	10	281.48	1.64	671
11	Smooth.	20	222.24	1.21	671



i	Method	ϵ	ν Avg	ν Unc	Time
0	Ref.	0	-6858.46	13.61	34840
1	DAAD	0	-6841.38	91.72	909
2	Smooth.	0.0001	-10967.9	11787.5	671
3	Smooth.	0.001	-5021.7	2433.64	671
4	Smooth.	0.01	-7055.03	939.43	671
5	Smooth.	0.1	-6954.56	296.30	671
6	Smooth.	1	-6642.21	91.88	671
7	Smooth.	2	-6035.03	60.68	671
8	Smooth.	3	-5176.24	44.24	671
9	Smooth.	5	-2999.04	24.12	671
10	Smooth.	10	2775.55	26.20	671
11	Smooth.	20	11160.1	67.73	671

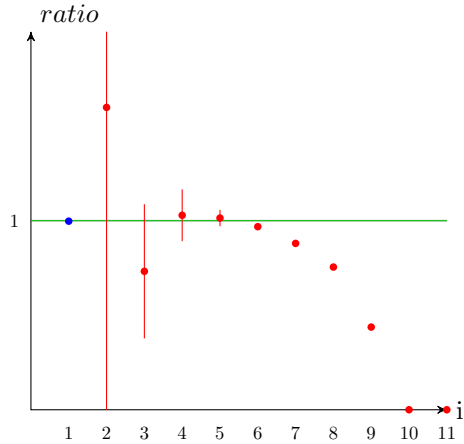
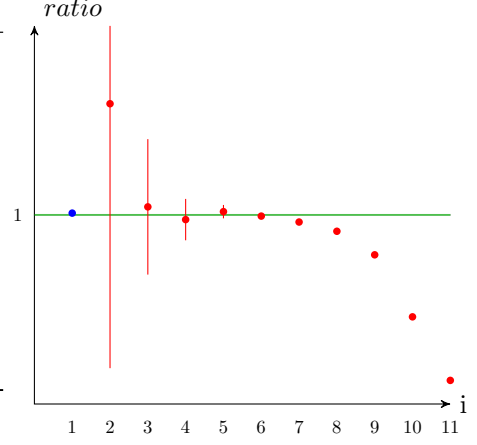


Figure 3.9: Delta and Vega sensitivity computed with DAAD and with ϵ -wide smoothing for the payoff $\phi = \prod_{i=1}^n H(A(T_i) - K_i)$ with $n = 20$, $T_i = 1 + i * 0.05$, strike $K_i = K = 100$, payment time $T_P = 2$. The green line (experiment $i = 0$) is a benchmark obtained by finite difference with a much higher number of scenarios. The only-price run time was 174 seconds.

i	Method	ϵ	δ Avg	δ Unc	Time
0	Ref.	0	445.29	1.10	34484
1	DAAD	0	449.58	5.02	1158
2	Smooth.	0.0001	707.21	622.83	478
3	Smooth.	0.001	464.31	159.51	478
4	Smooth.	0.01	434.32	48.73	478
5	Smooth.	0.1	453.05	15.78	478
6	Smooth.	1	442.55	5.02	478
7	Smooth.	2	428.58	3.58	478
8	Smooth.	3	406.84	2.93	478
9	Smooth.	5	351.36	2.23	478
10	Smooth.	10	205.28	1.32	478
11	Smooth.	20	55.60	0.49	478



i	Method	ϵ	ν Avg	ν Unc	Time
0	Ref.	0	-2338.33	7.99	34511
1	DAAD	0	-2326.81	45.64	1158
2	Smooth.	0.0001	-4641.85	6747.36	478
3	Smooth.	0.001	-1570.5	964.87	478
4	Smooth.	0.01	-2422.37	474.07	478
5	Smooth.	0.1	-2373.25	145.01	478
6	Smooth.	1	-2056.5	44.48	478
7	Smooth.	2	-1302.27	29.99	478
8	Smooth.	3	-302.84	24.97	478
9	Smooth.	5	1803.92	26.04	478
10	Smooth.	10	4278.93	35.23	478
11	Smooth.	20	2758.69	28.34	478

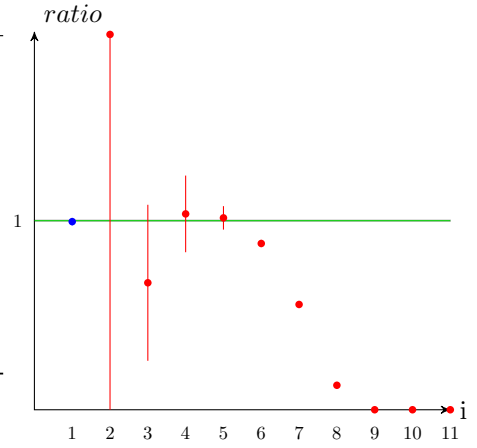
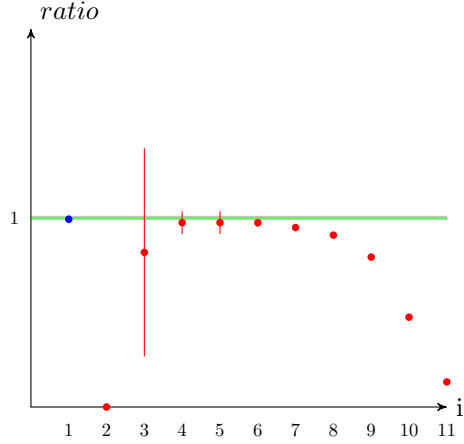


Figure 3.10: Delta and Vega sensitivity computed with DAAD and with ϵ -wide smoothing for the payoff $\phi = \prod_{i=1}^n H(A(T_i) - K_i)$ with $n = 40$, $T_i = i * 0.05$, strike $K_i = K = 100$, payment time $T_P = 2$. The green line (experiment $i = 0$) is a benchmark obtained by finite difference with a much higher number of scenarios. The only-price run time was 173 seconds.

i	Method	ϵ	δ Avg	δ Unc	Time
0	Ref.	0	152.07	1.32	34715
1	DAAD	0	151.01	2.92	1088
2	Smooth.	0.0001	0	0	526
3	Smooth.	0.001	124.34	83.76	526
4	Smooth.	0.01	148.25	9.18	526
5	Smooth.	0.1	148.3	9.21	526
6	Smooth.	1	148.19	3.00	526
7	Smooth.	2	144.34	2.16	526
8	Smooth.	3	138.23	1.78	526
9	Smooth.	5	120.59	1.38	526
10	Smooth.	10	72.23	0.82	526
11	Smooth.	20	20.16	0.30	526



i	Method	ϵ	ν Avg	ν Unc	Time
0	Ref.	0	4383.11	36.02	34741
1	DAAD	0	4493.7	173.44	1088
2	Smooth.	0.0001	0	0	526
3	Smooth.	0.001	2553.57	4573.02	526
4	Smooth.	0.01	2394.8	1504.14	526
5	Smooth.	0.1	4269.58	631.68	526
6	Smooth.	1	4446.84	83.76	526
7	Smooth.	2	4463.32	139.674	526
8	Smooth.	3	4550.58	113.61	526
9	Smooth.	5	4501.38	86.41	526
10	Smooth.	10	3643.32	55.95	526
11	Smooth.	20	1563.03	27.37	526

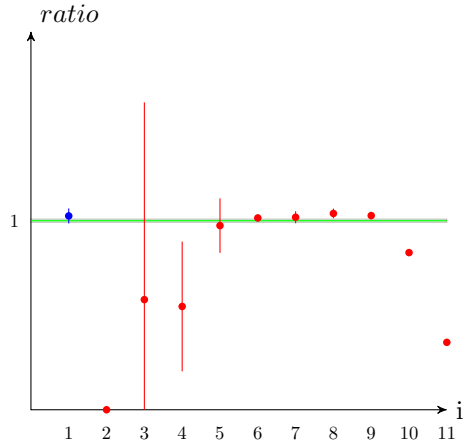


Figure 3.11: Delta and Vega sensitivity computed with DAAD and with ϵ -wide smoothing for the payoff $\phi = \prod_{i=1}^n H(A(T_i) - K_i)$ with $n = 40$, $T_i = i * 0.05$, strike $K_i = K = 100$ for $i \leq 20$, $K_i = K_{i-1} + 1$ for $i > 20$, payment time $T_P = 2$. The green line (experiment $i = 0$) is a benchmark obtained by finite difference with a much higher number of scenarios. The only-price run time was 174 seconds.

The results are in Figure 3.12, and confirm the qualitative remarks of Section 3.6.3.

3.6.5 Digital on two correlated assets

As a final multi-asset example, we study an instrument paying at time T_P an amount equal to 1 only if the average of the values of two correlated underlyings A_1 and A_2 at time T_P is above a certain strike. The discounted payoff can be written this way:

$$\phi = D(t, T_P) H(A_1(T_P) + A_2(T_P) - K). \quad (3.50)$$

The parameters of the dynamics of A_1 are the same as in the previous sections, while for the second underlying $A_2(0) = 50$, $\mu(t) = \mu = 2\%$, $\sigma(t) = \sigma = 15\%$. Instantaneous correlation between the two driving Wiener processes is put to 0.5.

The Delta and Vega sensitivities relative to the first asset are in Figure 3.13, and they reproduce the qualitative behaviour of the single-asset experiments.

3.7 Conclusions

We have presented a general technique to differentiate the Monte Carlo price of potentially-discontinuous payoffs. It generalises the popular pathwise method, and reduces exactly to it for differentiable payoffs.

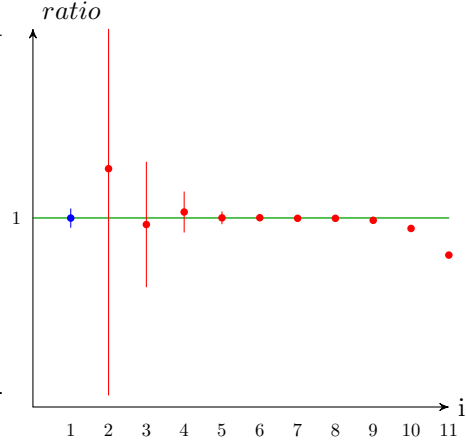
The implementation can be made very efficient especially for the computation of many sensitivities, thanks to the use in many points of adjoint differentiation. The parts of the algorithm which escape this boost can be reduced to a minimum at least when the risk factors are diffusions which are simulated with the Euler scheme, thanks to a formulation based on Brownian bridges.

An empirical analysis has shown promising performances in terms of accuracy, variance of the Monte Carlo estimator, and run times, when compared to the main benchmark at this level of generality, i.e. the pathwise differentiation of a smooth approximation of the payoff; whose choice has turned out to be quite delicate especially in presence of multiple discontinuities. The method proposed here, instead, does not introduce any bias and does not need the tuning of any arbitrary parameter.

Being the method new, the exposition and examples have focused on prototypical diffusive dynamics; but we believe that many ideas could be ported to different models for which the simulation step dominates the time spent evaluating the payoff. In this respect, some hope comes from Appendix A.2, where the Gaussianity hypothesis in the main formulas is relaxed. We leave out a more general analysis in this direction as a possible development.

Finally, we would like to note that the availability of a generalization of the pathwise method for discontinuous payoffs may be useful for the computation of sensitivities of order greater than one. Indeed, the standard pathwise method is very seldom applicable to second derivatives, because the payoff is almost never C^2 : the plainest call option is just piecewise C^1 . Hence one typically feels forced to design methods at least partially based on differentiation of the probability density (Pagès et al., 2016; Capriotti, 2015), with all the limitations of applicability discussed in the introduction about the likelihood-ratio method and the like. A detailed study of this kind of application is available in the next chapter.

i	Method	ϵ	δ Avg	δ Unc	Time
0	Ref.	0	481.56	1.14	34491
1	DAAD	0	481.20	24.44	498
2	Smooth.	0.0001	606.97	577.37	478
3	Smooth.	0.001	464.84	159.69	478
4	Smooth.	0.01	496.67	52.16	478
5	Smooth.	0.1	481.75	16.20	478
6	Smooth.	1	482.11	5.01	478
7	Smooth.	2	480.60	3.45	478
8	Smooth.	3	480.58	2.74	478
9	Smooth.	5	475.52	1.99	478
10	Smooth.	10	454.79	1.17	478
11	Smooth.	20	386.96	0.50	478



i	Method	ϵ	ν Avg	ν Unc	Time
0	Ref.	0	-6968.42	13.72	34476
1	DAAD	0	-6895	350.35	498
2	Smooth.	0.0001	-8369.09	7999.18	478
3	Smooth.	0.001	-6625.22	2281.36	478
4	Smooth.	0.01	-7109.97	748.17	478
5	Smooth.	0.1	-6917.09	233.11	478
6	Smooth.	1	-6913.76	73.48	478
7	Smooth.	2	-6878.1	53.57	478
8	Smooth.	3	-6878.57	46.68	478
9	Smooth.	5	-6804.87	43.03	478
10	Smooth.	10	-6371.72	46.64	478
11	Smooth.	20	-4714.09	53.19	478

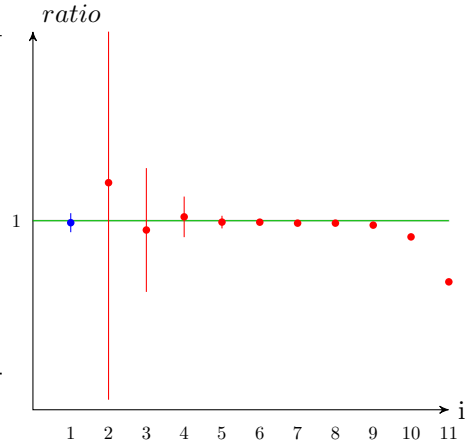


Figure 3.12: Delta and Vega sensitivity computed with DAAD and with ϵ -wide smoothing for the payoff $\phi = H(\frac{1}{n} \sum_{i=1}^n A(T_i) - K)$ with $n = 40$, $T_i = i * 0.05$, strike $K = 100$, payment time $T_P = 2$. The green line (experiment $i = 0$) is a benchmark obtained by finite difference with a much higher number of scenarios. The only-price run time was 170 seconds.

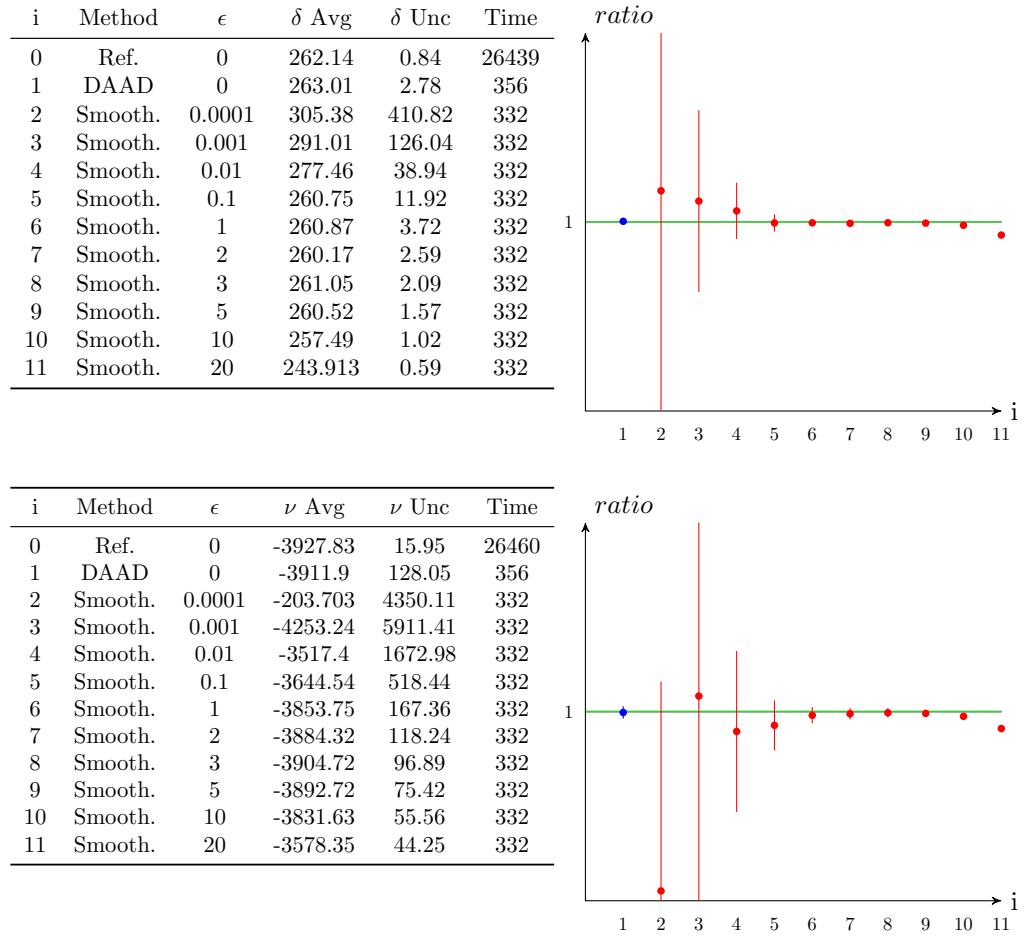


Figure 3.13: Delta and Vega sensitivity computed with DAAD and with ϵ -wide smoothing for the first underlying of the payoff $\phi = H(A_1(T_P) + A_2(T_P) - K)$ with $\rho_{A_1, A_2} = 0.5$, strike $K = 150$, payment time $T_P = 1$. The green line (experiment $i = 0$) is a benchmark obtained by finite difference with a much higher number of scenarios. The only-price run time was 132 seconds.

Chapter 4

Second order sensitivities in linear or constant time

In this chapter, we analyse and compare methods to compute the full set of second order sensitivities of a Monte Carlo price in a time which is at most $O(N \cdot T)$ where N is the number of inputs and T is the time required to compute the price. The new ones include the first algorithm which achieves a complexity $O(T)$ and has acceptable (in fact very low) statistical uncertainties at least in one relevant test case.

4.1 Introduction

As we saw throughout this thesis, in the active management of portfolios of financial products, the sensitivity of the net position to movements of the underlying risk factors is a most precious piece of information. Up to now, we have focused on first order Greeks, which are obviously the most important indicator for a trader to see whether his overall position is approximately market neutral, and which are often kept under strict control by dynamically adjusting the hedging positions as time passes. A direct consequence of this behaviour is that the day-to-day performance of the trading desk is at leading order driven by second order Greeks (also called Gammas). By the way, this is the discrete-time counterpart of the fact that in an idealized Black-Scholes world, the continuously rebalanced Delta hedging replication strategy generates an Ito component $\frac{1}{2}(\partial^2 P / \partial X_t^2)dt$ in the dynamics of total wealth.

All of the above implies that the knowledge of second order sensitivities is often crucial; which is even more true when the payoff depends on several and maybe correlated risk factors, whose co-movements impact the portfolio value through the mixed terms in the Hessian matrix (sometimes denoted Cross Gammas). This is the case for instance of basket products, indices, multi-asset-class hybrids, and portfolio-level non-linear valuation adjustments such as CVA and DVA (Basel Committee on Banking Supervision, 2015).

Unfortunately, these are exactly the settings in which the pricing exercise is computationally more demanding, as it often involves a multidimensional Monte Carlo simulation. This rules out as infeasible the simplest way to approximate the full

$N \times N$ pricing Hessian $\mathbf{H}$, i.e.

$$H_{ij} \approx \frac{P(\boldsymbol{\psi} + \mathbf{h}_i + \mathbf{h}_j) - P(\boldsymbol{\psi} + \mathbf{h}_i - \mathbf{h}_j) - P(\boldsymbol{\psi} - \mathbf{h}_i + \mathbf{h}_j) + P(\boldsymbol{\psi} - \mathbf{h}_i - \mathbf{h}_j)}{(1 + 3\delta_{ij})h^2},$$

where $\boldsymbol{\psi}$ is the vector of pricing inputs of interest, h is a small displacement and $\mathbf{h}_i := h\mathbf{e}_i$ is h times the i -th element of the canonical basis of $\mathbb{R}^N$: this method would involve $O(N^2)$ full revaluations.

In fact, the gradient ∇P can be computed in $O(T)$ where T is the time to compute P , thanks to adjoint algorithmic differentiation (AAD) which is more and more widespread for first order sensitivities among practitioners; see the previous chapter for details and references. This technique and its generalizations (e.g. Giles, 2009; Chan and Joshi, 2015; Daluiso and Facchinetti, 2018) coupled with the approximation

$$H_{ij} \approx \frac{\nabla_j P(\boldsymbol{\psi} + h\mathbf{e}_i) - \nabla_j P(\boldsymbol{\psi} - h\mathbf{e}_i)}{2h} \quad \text{or} \quad H_{ij} \approx \frac{\nabla_j P(\boldsymbol{\psi} + h\mathbf{e}_i) - \nabla_j P(\boldsymbol{\psi})}{h},$$

would already give an estimator in $O(N \cdot T)$, but this is in practice quite unsatisfactory, since for small h the result is very noisy, while for moderate h it is too much biased.

The latter fact stimulated research of alternative algorithms which could retain the $O(N \cdot T)$ cost while being unbiased and having less Monte Carlo variance. We point out in particular the significantly different contributions Capriotti (2015), Pagès et al. (2016) and Joshi and Zhu (2016), which seem unaware of one another and present numerical evidences which are difficult to compare.

Our first contribution is therefore a comparative analysis of the three above proposals, to see in which settings they can be applied. In fact, while not delving into the details of the estimators which are already well covered in the original papers, we sometimes go slightly beyond the authors: namely, by underlying a subtlety needed to correctly implement Capriotti (2015), and by developing a multi-fixing generalization of Pagès et al. (2016).

As we will see, in some situations the rich set of alternatives outlined above shrinks considerably: in particular, payoffs directly depending on the non random initial state or on static parameters (e.g. forwards and deterministic interest rates), and factorial models in general where the number of random drivers is smaller than the number of underlyings, are usually out of scope. This motivated the development of the second contribution of this chapter, namely a fairly general-purpose method based on second order pathwise differentiation of the payoff in distributional sense.

Finally, most of the above methods run in $O(N \cdot T)$ time. This is a theoretical upper bound for general computer functions (see e.g. Griewank and Walther, 2008), but can the special structure of a pricing problem with diffusive underlyings change the verdict? To our knowledge, the only partial positive answer is provided by the already cited Joshi and Zhu (2016) in the case in which the simulation can be performed keeping at each time step only a low dimensional state variable.

In this respect, our third and last contribution is a positive answer to the above question in the complementary setting in which the Brownian driver is high dimensional (precisely, has dimensionality at least equal to the number of underlyings). Our solution is based on a relation linking a suitable first order sensitivity to $\partial^2 P / \partial \mathbf{X}_0^2$ via the pricing PDE, and yields the latter in $O(T)$ time by an easily implementable wrapper of any $O(T)$ first order estimator. In fact, also LRMAAD from Capriotti (2015) has constant cost and almost the same perimeter of applicability, but from the numerical analysis it turns out to have too large Monte Carlo uncertainties to be an

option in practice; this is sometimes true also for the new estimator, but we will show at least one relevant example in which our new constant-cost algorithm significantly outperforms all its linear-cost competitors.

The rest of the chapter is structured as follows. Section 4.2 states the problem in the diffusive setting in which all our analyses will be performed; Section 4.3 reviews the methods proposed in the literature; Section 4.4 introduces the new methods; Section 4.5 compares the empirical performances and Monte Carlo uncertainties on several test cases; finally, Section 4.6 summarizes the main findings and concludes.

4.2 Setting

As in Section 3.5, we consider a description of the market where the risk factors $\mathbf{X}_t$ are modelled as an $\mathbb{R}^{N_X}$ -valued diffusion driven by an N_W -dimensional Wiener process $\mathbf{W}_t$ under a fixed probability measure $\mathbb{P}$:

$$d\mathbf{X}_t = \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{X}_t, t)dt + \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}_t, t)d\mathbf{W}_t, \quad t \geq 0, \quad (4.1)$$

where $\boldsymbol{\theta} \in \mathbb{R}^{N_\theta}$ is a vector of parameters. A price P of interest is expressed as an expected value

$$P = \mathbb{E}[g(\mathbf{X}_{T_1}, \dots, \mathbf{X}_{T_M}, \boldsymbol{\theta})] =: \mathbb{E}[g(\mathbf{X}_T, \boldsymbol{\theta})], \quad (4.2)$$

and is estimated by Monte Carlo simulation; to be concrete, we suppose that the simulation is performed by Euler stepping on a time grid $\{S_0, \dots, S_L\} = \mathbf{S} \supseteq \mathbf{T}$:

$$\begin{aligned} \mathbf{X}_{S_{i+1}} &= \mathbf{X}_{S_i} + \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{X}_{S_i}, S_i) \Delta_i S + \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}_{S_i}, S_i) \sqrt{\Delta_i S} \cdot \mathbf{Z}^{(i)}, \\ \Delta_i S &:= S_{i+1} - S_i, \quad \mathbf{Z}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \end{aligned} \quad (4.3)$$

although one could imagine more or less straightforward generalizations of the methods described below for more general schemes.

When the payoff does not depend on $\boldsymbol{\theta}$ directly, we will write

$$g(\mathbf{X}_T, \boldsymbol{\theta}) =: p(\mathbf{X}_T); \quad (4.4)$$

we will call “autonomous” this simpler case. Note that the price P still depends on $\boldsymbol{\theta}$ through the dynamics (4.1).

We aim at computing the full Gamma matrix

$$\boldsymbol{\Gamma} := \frac{\partial^2 P}{\partial \mathbf{X}_0^2},$$

and possibly the full Hessian

$$\mathbf{H} := \frac{\partial^2 P}{\partial \boldsymbol{\psi}^2}, \quad \boldsymbol{\psi} := (\mathbf{X}_0, \boldsymbol{\theta}),$$

in acceptable time when N_X and N_θ are large. Note that a naive approach based on finite differences would involve

$$O\left((N_X + N_\theta)^2\right) =: O(N^2)$$

re-evaluations of the price P with displaced values of $\boldsymbol{\theta}$ and/or $\mathbf{X}_0$.

Name	1st order	2nd order	Section	Reference
FDIFF2	finite difference	finite difference	4.1	N.A.
FDPW	pathwise AD	finite difference	4.3.1	N.A.
LRMAAD	pathwise AD	likelihood ratio	4.3.2	Capriotti (2015)
AADLRM	likelihood ratio	pathwise AD	4.3.2	Capriotti (2015)
VAD	vibrato	pathwise AD	4.3.3	Pages et al. (2016)
VFD	vibrato	finite difference	4.3.3	N.A.
HOPP	change var + AD	pathwise AD	4.3.4	Joshi and Zhu (2016)

Table 4.1: Legacy methods to compute second order sensitivities: estimators used for first and second order differentiation, references. Here as elsewhere in this thesis, AD stands for algorithmic differentiation.

4.3 Review of existing algorithms

In this section, we review from an algorithmic point of view the main methods based on the existing literature whose theoretical complexity is $O(N \cdot T)$, with a focus on the perimeter of applicability within the setting of Section 4.2. For quick reference, we gather in Table 4.1 the essential traits of all these algorithms.

4.3.1 Finite differences of pathwise adjoint

Before describing more complex methods, let us mention that the computation of finite differences of the pathwise adjoint estimator is already a linear cost method. Indeed, first order pathwise adjoints cost at most $4T$; repeating their computation N times for small directional displacements of ψ , one trivially gets an estimator for the Hessian, which will be denoted by the acronym FDPW.

However, we expect poor variance properties of this naive method, as finite differences have high Monte Carlo uncertainties when the integrand is discontinuous, as is the case for the gradient of typical payoffs. Therefore, we explore in the following sections analytical alternatives leveraging on the structure of the problem.

4.3.2 Likelihood ratio and adjoint differentiation

The idea of Capriotti (2015) is to combine pathwise differentiation with the likelihood ratio method. Since the latter can handle only the autonomous case, the same limitation applies to the resulting second order methods; moreover, as is clear from the matrix inversion in equation (4.6) here below, one needs that $N_X \leq N_W$ with $\Sigma(\theta, \mathbf{X}_T, T)$ almost surely full rank. As we believe that a crucial detail in the correct interpretation of the final formula may be easily overlooked, we repeat the derivation briefly.

When adjoint algorithmic differentiation is applied first, one wants to compute

$$\frac{\partial}{\partial \psi} \mathbb{E} \left[\frac{\partial p}{\partial \mathbf{X}_T}(\mathbf{X}_T) \frac{\partial \mathbf{X}_T}{\partial \psi}(\mathbf{Z}, \psi) \right]$$

by likelihood ratio. In order to do so, one *must* rewrite the argument of $\mathbb{E}$ so that the only random variable appearing is $\mathbf{X}_S$, because the method relies on writing the expectation as an integral against the density of $\mathbf{X}_S$. Fortunately, $\mathbf{Z}^{(i)}$ is readily

expressed as a function of $\mathbf{X}_S$ and $\boldsymbol{\theta}$ as

$$z^{(i)}(\mathbf{X}_S, \boldsymbol{\theta}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}_{S_i}, S_i)^{-1} (\Delta_i S)^{-\frac{1}{2}} (\mathbf{X}_{S_{i+1}} - \mathbf{X}_{S_i} - \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{X}_{S_i}, S_i) \Delta_i S).$$

Now a straightforward generalization of the likelihood ratio method to the case in which the integrand has explicit dependence on the differentiation parameter leads to the following estimator of the Hessian $\mathbf{H}$:

$$\left(\frac{\partial p}{\partial \mathbf{X}_T}(\mathbf{X}_T) \frac{\partial \mathbf{X}_T}{\partial \boldsymbol{\psi}}(\mathbf{Z}, \boldsymbol{\psi}) \right)^\top \boldsymbol{\Omega}_\psi + \frac{\partial}{\partial \boldsymbol{\psi}} \left[\frac{\partial p}{\partial \mathbf{X}_T}(\mathbf{X}_T) \frac{\partial \mathbf{X}_T}{\partial \boldsymbol{\psi}}(z(\mathbf{X}_S, \boldsymbol{\theta}), \boldsymbol{\psi}) \right], \quad (4.5)$$

where the $\boldsymbol{\Omega}_\psi$ is the customary likelihood ratio weight capturing the component of the derivative due to dependence of the distribution of $\mathbf{X}_S$ on $\boldsymbol{\psi}$ by differentiation of the log-density, while the future path $\mathbf{X}_{S \setminus \{0\}}$ in the square bracket is a constant for the purpose of differentiation. Note that in general the second addend cannot be replaced by the simpler

$$\frac{\partial p}{\partial \mathbf{X}_T}(\mathbf{X}_T) \frac{\partial}{\partial \boldsymbol{\psi}} \left[\frac{\partial \mathbf{X}_T}{\partial \boldsymbol{\psi}}(\mathbf{Z}, \boldsymbol{\psi}) \right]$$

which seems implicit in the implementation described by Capriotti (2015).

Therefore, the algorithm is as follows:

Algorithm 4.1 (LRMAAD).

1. Compute simultaneously all $\bar{\mathbf{X}}_{T_i} = \frac{\partial p}{\partial \mathbf{X}_{T_i}}$ with AAD.
2. Compute the derivatives of the Gaussian transitions' log-densities as

$$\begin{aligned} \bar{\boldsymbol{\mu}}^{(i)} &= \left(\mathbf{X}_{S_{i+1}} - \boldsymbol{\mu}^{(i)} \right)^\top V_i^{-1}, \\ \bar{\boldsymbol{\Sigma}}^{(i)} &= \left(\boldsymbol{\Sigma}^{(i)} \right)^\top \left[\left(\bar{\boldsymbol{\mu}}^{(i)} \right)^\top \left(\bar{\boldsymbol{\mu}}^{(i)} \right) - V_i^{-1} \right], \end{aligned} \quad (4.6)$$

where the mean, standard deviation and covariance matrices are defined as

$$\begin{aligned} \boldsymbol{\mu}^{(i)} &= \mathbf{X}_{S_i} + \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{X}_{S_i}, S_i) \Delta_i S, \\ \boldsymbol{\Sigma}^{(i)} &= \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}_{S_i}, S_i) \sqrt{\Delta_i S}, \quad V_i = \left(\boldsymbol{\Sigma}^{(i)} \right) \left(\boldsymbol{\Sigma}^{(i)} \right)^\top. \end{aligned}$$

3. Propagate them backward with AAD to $\boldsymbol{\psi}$ obtaining likelihood ratio weights Ω_j for $j = 1, \dots, N$.
4. Propagate $\bar{\mathbf{X}}_{T_i}$ backward through simulation obtaining $\bar{\boldsymbol{\psi}} = \sum_{i=1}^M \bar{\mathbf{X}}_{T_i} \cdot \frac{\partial \mathbf{X}_{T_i}}{\partial \boldsymbol{\psi}}$.
5. Apply algorithmic differentiation or finite differences to the previous step, where $\bar{\mathbf{X}}_{T_i}$ should be considered as constants while $\mathbf{Z}^{(i)}$ should be *recomputed* as

$$\left(\boldsymbol{\Sigma}^{(i)} \right)^{-1} \left(\mathbf{X}_{S_{i+1}} - \boldsymbol{\mu}^{(i)} \right).$$

Call $\mathbf{J}$ the resulting Jacobian with respect to $\boldsymbol{\psi}$.

6. The estimator for H_{ij} is $J_{ij} + \bar{\psi}_i \Omega_j$.

Only step 5 has a cost linearly growing with N . This is unavoidable in general, since computing a Jacobian $\mathbf{J}$ multiplies the run time of a function by $O(N)$, regardless of whether one uses AAD as suggested in the original article, or a simpler forward differentiation (Griewank and Walther, 2008), or even finite differences.

However, if one wants to determine only $\mathbf{\Gamma}$, then this term can be computed very efficiently. Indeed, the step under analysis computes the Jacobian of the second addend of (4.5), which for $\psi = \mathbf{X}_0$ reduces to

$$\frac{\partial}{\partial \mathbf{X}_0} \left[\frac{\partial p}{\partial \mathbf{X}_T}(\mathbf{X}_T) \frac{\partial \mathbf{X}_T}{\partial \mathbf{X}_0}(z(\mathbf{X}_S), \mathbf{X}_0) \right];$$

now, one can easily note that if the quantity in square brackets is computed by AAD, then in the implementation it takes the form

$$\bar{\mathbf{X}}_{S_1} \cdot \frac{\partial \mathbf{X}_{S_1}}{\partial \mathbf{X}_0}(z(\mathbf{X}_{S_1}, \mathbf{X}_0), \mathbf{X}_0)$$

for a suitable random variable $\bar{\mathbf{X}}_{S_1}$ *not depending on* $\mathbf{X}_0$. Once this is remarked, differentiation of this term is so fast to become practically negligible, and the complexity of the algorithm is dominated by the first term in (4.5), computed in $O(T)$ time. Note that even there, the likelihood ratio weight $\Omega_{\mathbf{X}_0}$ has non-trivial contribution from (4.6) only for the first step $i = 0$: this makes the overall runtime essentially comparable to that of computing the first order pathwise estimator

$$\frac{\partial p}{\partial \mathbf{X}_T}(\mathbf{X}_T) \frac{\partial \mathbf{X}_T}{\partial \mathbf{X}_0}(\mathbf{Z}, \mathbf{X}_0).$$

Instead, if one differentiates pathwise the first order likelihood ratio estimator, the following is obtained:

Algorithm 4.2 (AADLRM). Implement Algorithm 4.1 replacing the steps:

5. Apply algorithmic differentiation (forward or backward) or finite differences to the computation of Ω (i.e. of the simulation scheme and of steps 2 and 3), obtaining its Jacobian $\mathbf{J}$ with respect to ψ .
6. The estimator for H_{ij} is $J_{ij}p(\mathbf{X}_T) + \bar{\psi}_j \Omega_i$.

Again the Jacobian step 5 is a potential $O(N)$ factor on run times. The original paper notes that this second method is empirically slower on a simple example, maybe because of the above $\mathbf{\Gamma}$ -specific speedup; however, we feel that a comparison of Monte Carlo variances is needed before discarding AADLRM as less performing.

4.3.3 Vibrato and adjoint differentiation

Pagès et al. (2016) apply automatic differentiation to a different first order estimator called Vibrato Monte Carlo, which has much better sample variance properties especially after the introduction of a trick based on the antithetic transform.¹ (They also consider the possibility of applying Vibrato twice, but they find no relevant difference.) As in the previous section, the analysis works only for the autonomous case and under the hypothesis $N_X \leq N_W$ with $\Sigma(\boldsymbol{\theta}, \mathbf{X}_T, T)$ almost surely full rank.

The algorithm to compute the estimator, here generalized to $M > 1$ fixing times in the spirit of what Giles (2009) mentions for first order Vibrato, is as follows:

¹Note that the antithetic idea may also be beneficial for the likelihood ratio method itself, as first remarked in Capriotti (2008).

Algorithm 4.3 (VAD).²

1. For each fixing time $T_i = S_{t(i)+1}$, compute the derivatives of the (approximate) Gaussian log-density of $\mathbf{X}_{T_i}$ given its neighbours in the simulation grid $\mathbf{S}$: i.e. apply (4.6) where

$$\begin{aligned}
\boldsymbol{\mu}^{(N)} &= \mathbf{X}_{S_{t(N)}} + \boldsymbol{\mu}(\boldsymbol{\theta}, \mathbf{X}_{S_{t(N)}}, S_{t(N)}) \Delta_{t(N)} S, \\
\boldsymbol{\Sigma}^{(N)} &= \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}_{S_{t(N)}}, S_{t(N)}) \sqrt{\Delta_{t(N)} S}, \\
\boldsymbol{\mu}^{(i)} &= (1 - \lambda_{t(i)}) \mathbf{X}_{S_{t(i)}} + \lambda_{t(i)} \mathbf{X}_{S_{t(i)+2}} \quad \text{for } i < N, \\
\boldsymbol{\Sigma}^{(i)} &= \sqrt{(\Delta_{t(i)} S) (1 - \lambda_{t(i)})} \boldsymbol{\Sigma}(\boldsymbol{\theta}, \mathbf{X}_{S_{t(i)}}, S_{t(i)}) \quad \text{for } i < N, \\
\lambda_t &:= (\Delta_t S + \Delta_{t+1} S)^{-1} \Delta_t S \quad \text{and} \quad V_i := \left(\boldsymbol{\Sigma}^{(i)} \right) \left(\boldsymbol{\Sigma}^{(i)} \right)^\top \quad \text{for } i \leq N.
\end{aligned} \tag{4.7}$$

2. For each fixing time $T_i = S_{t(i)+1}$, draw $\mathbf{X}_{T_i}$ from the conditional law:

$$\mathbf{X}_{T_i} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\Sigma}^{(i)} \mathbf{W}^{(i)}, \quad \mathbf{W}^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_W}) \tag{4.8}$$

3. Compute the antithetic variates

$$\begin{aligned}
p_i^- &:= p\left(X_{T_1}, \dots, \boldsymbol{\mu}^{(i)} - \boldsymbol{\Sigma}^{(i)} \mathbf{W}^{(i)}, \dots, X_{T_M}\right), \\
p_i^0 &:= p\left(X_{T_1}, \dots, \boldsymbol{\mu}^{(i)}, \dots, X_{T_M}\right).
\end{aligned}$$

4. Multiply $\bar{\boldsymbol{\mu}}^{(i)}$ by $\frac{p - p_i^-}{2}$ and $\bar{\boldsymbol{\Sigma}}^{(i)}$ by $\frac{p + p_i^-}{2} - p_i^0$.
5. Repeat steps 2 to 4 for N_{in} independent draws of the random numbers $\mathbf{W}$ and average the resulting $\bar{\boldsymbol{\mu}}^{(i)}$, $\bar{\boldsymbol{\Sigma}}^{(i)}$.
6. Propagate backwards $\bar{\boldsymbol{\mu}}^{(i)}$, $\bar{\boldsymbol{\Sigma}}^{(i)}$ through the adjoint of (4.7) and of the Euler scheme, obtaining sensitivities $\bar{\boldsymbol{\psi}}$.
7. Apply automatic differentiation (forward or backward) to the whole algorithm.

Note that the last step multiplies by $O(N)$ the cost of the whole algorithm. As for the methods of the previous section, if the payoff is discontinuous then the method cannot be applied as is; however, in this case we can conceive to perform the last step with a finite difference with non-small displacement: we will denote this variant as VFD.

Finally, the choice of N_{in} would need investigation, since Pagès et al. (2016) suggest a small N_{in} (say 1 or 2), while Giles (2009) finds that at least for first order sensitivities a good value of N_{in} from the variance point of view should be of the order of the square root of the outer Monte Carlo scenarios. However, when coupled with the antithetic technique as generalized above for $M > 1$ fixings, the latter advice would result in a complexity per outer scenario which is much higher than in the original pricing, and this may be a sufficient concern in applications to give up the improvement in asymptotic uncertainties.

²The algorithms in the previous section are called LRMAAD when LRM is applied *after* AAD, and AADLRM otherwise; while here, somewhat confusingly, VAD is the method which applies V(ibrato) *before* AD. We keep the acronyms chosen by the original authors nonetheless.

4.3.4 Optimal partial proxy for Hessians

In real payoffs, direct pathwise second order differentiation is not possible only because of a few discontinuity points of the gradient or of the payoff. Exploiting this fact, Joshi and Zhu (2016) manage to change the integrand without changing the integral, so that one can differentiate algorithmically twice the modified payoff. More precisely, they rewrite the expected value as a function of a multivariate uniform $\mathbf{U}$,

$$P = \mathbb{E}[g(\mathbf{X}_T, \boldsymbol{\theta})] =: \mathbb{E}[g(\mathbf{x}_T(\mathbf{Z}, \boldsymbol{\psi}), \boldsymbol{\theta})] =: \mathbb{E}[g(\mathbf{x}_T(\mathcal{N}^{-1}(\mathbf{U}), \boldsymbol{\psi}), \boldsymbol{\theta})]$$

and then perform a change of variables $\mathbf{V} = \mathbf{v}(\mathbf{U}, \boldsymbol{\psi})$, designed in such a way that while moving $\boldsymbol{\psi}$, the path generated by $\mathbf{v}(\mathbf{U}, \boldsymbol{\psi})$ does not cross the discontinuities. After the change of variables,

$$P = \mathbb{E}[g(\mathbf{x}_T(\mathcal{N}^{-1}(\mathbf{v}(\mathbf{U}, \boldsymbol{\psi})), \boldsymbol{\psi}), \boldsymbol{\theta}) \cdot \omega(\mathbf{U}, \boldsymbol{\psi})]$$

for a suitable weight ω ; the mapping $\mathbf{v}$ should ensure that if the payoff is smooth except when

$$U_{N_W}^{(t(i))} = a_j^{(i)}(\boldsymbol{\psi}, U^{(1)}, \dots, U^{(t(i)-1)}, U_1^{(t(i))}, \dots, U_{N_W-1}^{(t(i))}) \quad \text{for some } j,$$

where $\{T_{t(i)+1}\}_{i=1, \dots, h}$ are the maturities of the payoff's discontinuities, then³

$$U_{N_W}^{(t(i))} \in \left[a_j^{(i)}(\boldsymbol{\psi}, \mathbf{U}), a_{j+1}^{(i)}(\boldsymbol{\psi}, \mathbf{U}) \right) \iff \forall \boldsymbol{\psi}, V_{N_W}^{(t(i))} \in \left[a_j^{(i)}(\boldsymbol{\psi}, \mathbf{V}), a_{j+1}^{(i)}(\boldsymbol{\psi}, \mathbf{V}) \right).$$

The simplest way to build such $\mathbf{v}$ is to make $\mathbf{v}^{(t)}$ depend only on $(\mathbf{U}^{(s)})_{s \leq t}$ and keep it equal to the identity mapping except on coordinates $(v_{N_W}^{(t(i))})_{i \geq 1}$; this way the change-of-variables weight is

$$\omega(\mathbf{U}, \boldsymbol{\psi}) = \prod_{i \geq 1} \frac{\partial v_{N_W}^{(t(i))}}{\partial U_{N_W}^{(t(i))}}.$$

The remaining degrees of freedom are fixed requiring that $\mathbf{v}(\mathbf{U}, \boldsymbol{\psi}_0) = \mathbf{U}$ for a reference value $\boldsymbol{\psi}_0$ which is the point where ∇P is to be estimated, and then minimizing the conditional variance of $\partial \omega / \partial \boldsymbol{\psi}$: it turns out that the best $\mathbf{v}$ in this sense is obtained when $V_{N_W}^{(t(i))}$ linearly interpolates between $a_j^{(i)}$ and $a_{j+1}^{(i)}$.⁴

Joshi and Zhu (2016) proceed suggesting an implementation based on a smart backpropagation of second order derivatives (Joshi and Yang, 2011), which should improve the computational burden when there are significantly fewer state variables in the stochastic process than parameters in the model. We do not pursue this level of optimization in our implementation, which aims at being generic; the analysis of the applicability and efficacy of such model-dependent and payoff-dependent trick on the different competitors described in this chapter could be the subject of future research.⁵ For the moment being, we sketch a less *ad hoc* algorithm to compute the estimator:

³At least to second order, as the original article points out.

⁴In fact, between their second order Taylor expansion in $\boldsymbol{\psi}$ around $\boldsymbol{\psi}_0$.

⁵Note that the optimization would not apply anyway to the numerical tests of this chapter, whose validity is therefore unaffected by the simplification.

Name	1st order	2nd order	Section
FDDAAD	distributional AD	finite differences	4.4.2
DAAD2	(distributional) AD	distributional AD	4.4.2
FT	any	none	4.4.3

Table 4.2: New methods to compute second order sensitivities: estimators used for first and second order differentiation, references. See also Table 4.1 for legacy methods.

Algorithm 4.4 (HOPP).

1. During simulation, compute the critical values $a_j^{(i)}$, the transformed drivers $\mathbf{V}$ and the weight ω . (In fact we know *a priori* that $\mathbf{V} = \mathbf{U}$ and $\omega = 1$ at $\boldsymbol{\psi} = \boldsymbol{\psi}_0$, but the instructions computing $\mathbf{V}$ and ω should be taken into account when in the sequel this step will be differentiated.)
2. Multiply the payoff g by the weight ω (with the same remark about $\omega = 1$).
3. Differentiate algorithmically twice the previous two steps, inclusive of simulation, with respect to $\boldsymbol{\psi}$ for fixed $\boldsymbol{\psi}_0$.

Step 3 is an $O(N)$ multiplier on the execution time of the first two steps, which in turn is typically little more than that of a plain Monte Carlo run, the overhead of computing ω being slightly significant only when the discontinuities are very frequent as in barrier options.

4.4 New methods

In this section we move to new proposals which try to overcome some limitations of the above methods. Table 4.2 completes with the contents of this section the summary which was provided in Table 4.1 for the algorithms of Section 4.3.

4.4.1 Second order differentiation by conditioning

We isolate in this small section a simple mathematical property, which will be used to derive the new second order method of Section 4.4.2, but which is meaningful *per se*. The general question is whether the ability to compute efficiently an estimator of the Hessian of a *conditional* expectation can enable the fast computation of the Hessian of the unconditional expectation.

More specifically, we take a set of times $\tilde{\mathbf{T}} \subseteq \mathbf{S}$, and let

$$\varepsilon(\mathbf{X}_{\tilde{\mathbf{T}}}, \boldsymbol{\theta}) = \mathbb{E}[g(\mathbf{X}_{\mathbf{T}}, \boldsymbol{\theta}) | \mathbf{X}_{\tilde{\mathbf{T}}}] .$$

Then one can compute the desired Hessian as $\mathbb{E}\left[\frac{\partial^2}{\partial \boldsymbol{\psi}^2} \varepsilon(\mathbf{X}_{\tilde{\mathbf{T}}}, \boldsymbol{\theta})\right]$, i.e.

$$\mathbb{E}\left[\left(\frac{\partial(\mathbf{X}_{\tilde{\mathbf{T}}}, \boldsymbol{\theta})}{\partial \boldsymbol{\psi}}\right)^\top \frac{\partial^2 \varepsilon}{\partial (\mathbf{X}_{\tilde{\mathbf{T}}}, \boldsymbol{\theta})^2} \left(\frac{\partial(\mathbf{X}_{\tilde{\mathbf{T}}}, \boldsymbol{\theta})}{\partial \boldsymbol{\psi}}\right) + \frac{\partial \varepsilon}{\partial \mathbf{X}_{\tilde{\mathbf{T}}}} \left(\frac{\partial^2 \mathbf{X}_{\tilde{\mathbf{T}}}}{\partial \boldsymbol{\psi}^2}\right)\right], \quad (4.9)$$

where by the tower rule, one can substitute the terms

$$\frac{\partial \varepsilon}{\partial \mathbf{X}_{\tilde{\mathbf{T}}}}, \frac{\partial^2 \varepsilon}{\partial (\mathbf{X}_{\tilde{\mathbf{T}}}, \boldsymbol{\theta})^2}$$

with $\mathbf{X}_{\tilde{T}}$ -conditional sample estimators. We note that the second addend of (4.9) is readily computed in linear time by differentiation of the function

$$\psi \mapsto \bar{\mathbf{X}}_{\tilde{T}} \cdot \frac{\partial \mathbf{X}_{\tilde{T}}}{\partial \psi}, \quad \text{with} \quad \bar{\mathbf{X}}_{\tilde{T}} \text{ any estimator of } \frac{\partial \varepsilon}{\partial \mathbf{X}_{\tilde{T}}},$$

which function is easily computed by pathwise AAD; the Jacobians in the first addend of (4.9) are similarly not a problem, and so we just need an estimator of the Hessian of ε , as desired.

4.4.2 Distributional algorithmic differentiation

The methods of Section 4.3 are devised to avoid at least in part the need of second order differentiation of the payoff, because its first derivative is typically discontinuous: a vanilla call already has

$$\frac{d}{dX}(X - K)^+ = I_{X > K}.$$

However, in Chapter 3 we presented a generalization of the first order pathwise method which works for discontinuous payoffs. Then on the one hand, by finite differences one may immediately form a simple biased second order estimator for discontinuous payoffs (FDDAAD); on the other hand, a new unbiased method for both continuous and discontinuous payoffs working also in the non-autonomous case can be introduced as follows.

As in most applications, suppose that there exist smooth scalar functions

$$f_i(\mathbf{X}_T, \boldsymbol{\theta}), g_{\mathbf{a}}(\mathbf{X}_T, \boldsymbol{\theta}) \quad \text{for} \quad i \in 1, \dots, h, \mathbf{a} \in \{-1, +1\}^h,$$

such that g is naturally represented as

$$g(\mathbf{X}_T, \boldsymbol{\theta}) = g_{\mathbf{a}}(\mathbf{X}_T, \boldsymbol{\theta}) \quad \text{on } \mathcal{X}_{\mathbf{a}} := \{(\mathbf{X}_T, \boldsymbol{\theta}) : a_i \cdot f_i(\mathbf{X}_T, \boldsymbol{\theta}) \geq 0 \forall i = 1, \dots, h\}.$$

We call $M_i = S_{t(i)+1}$ the last time in T such that f_i depends on $\mathbf{X}_{M_i}$; we suppose that we can find $k(i)$ such that the $k(i)$ -th component of

$$\frac{\partial f_i}{\partial \mathbf{Z}^{(t(i))}} = \frac{\partial f_i}{\partial \mathbf{X}_{M_i}} \boldsymbol{\Sigma}^{(t(i))}$$

is almost-surely different from zero.

Now we exploit the conditioning idea of (4.9) using as $\tilde{T}$ the (often quite small) set obtained substituting in T each M_i with its first neighbours in the simulation grid S . Recall that $\mathbf{X}_{M_i}$ can be simulated conditionally on its neighbours by Brownian bridge interpolation (4.8). With this choice of $\tilde{T}$, we can estimate the Hessian of ε by applying first pathwise differentiation to its definition (with distributional corrections if the payoff itself is discontinuous), and then distributional algorithmic differentiation to the result.

A brief derivation of the full estimator for possibly discontinuous payoffs is available in Appendix A.4; here, for the sake of clarity, we detail the algorithm under the hypothesis that only the gradient is discontinuous.

Call $\tilde{f}_i, \tilde{g}$ the functions computing f_i, g from $\mathbf{X}_{\tilde{T}}$, including conditional simulation of $\mathbf{X}_{M_i}$. Then the algorithm on each Monte Carlo scenario is as follows:

Algorithm 4.5. [DAAD2]

1. Apply pathwise adjoint differentiation to the composition of g with the conditional simulation of the $\mathbf{X}_{M_i}$, getting estimators $\boldsymbol{\theta}, \bar{\mathbf{X}}_{\tilde{T}}$.

2. Differentiate again the previous step ignoring discontinuities, getting a matrix $\mathbf{D}_0$ representing the smooth part of the $\mathbf{X}_{\tilde{T}}$ -conditional sample estimator for the Hessian of ε .
3. For each $i = 1, \dots, h$ compute the values ω which substituted into $W_{k(i)}^{(i)}$ make a $\tilde{\mathbf{W}}^{(i)}$ such that f_i equals zero, and the corresponding perturbed version of $\mathbf{X}_T$, which we denote by $\tilde{\mathbf{X}}_T$. Then for each such ω :
 - (a) There should be exactly two $\mathbf{a} \in \{-1, +1\}^h$ such that $\tilde{\mathbf{X}}_T \in \mathcal{X}_{\mathbf{a}}$, corresponding to $a_i = \pm 1$: call them $\mathbf{a}_{\pm}$.
 - (b) Use AAD to compute in the perturbed scenario the gradients

$$\frac{\partial \tilde{f}_i}{\partial \mathbf{W}^{(i)}}, \frac{\partial \tilde{f}_i}{\partial \mathbf{X}_{\tilde{T}}}, \frac{\partial \tilde{f}_i}{\partial \boldsymbol{\theta}}, \frac{\partial \tilde{g}_{\mathbf{a}_{\pm}}}{\partial \mathbf{X}_{\tilde{T}}}, \frac{\partial \tilde{g}_{\mathbf{a}_{\pm}}}{\partial \boldsymbol{\theta}}.$$

- (c) Increment the discontinuity's contribution to $\frac{\partial^2 \varepsilon}{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})^2}$ as follows:

$$\mathbf{D}_i += \frac{e^{-\frac{\omega^2}{2}}}{\sqrt{2\pi}} \left| \frac{\partial \tilde{f}_i}{\partial W_{k(i)}^{(i)}} \right|^{-1} \left(\frac{\partial \tilde{f}_i}{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})} \right)^{\top} \left[\frac{\partial g_{\mathbf{a}_{+}}}{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})} - \frac{\partial g_{\mathbf{a}_{-}}}{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})} \right]. \quad (4.10)$$

4. Compute the Jacobian $\frac{\partial \mathbf{X}_{\tilde{T}}}{\partial \boldsymbol{\psi}}$, which is the only unknown part of $\frac{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})}{\partial \boldsymbol{\psi}}$.
5. Compute by adjoint differentiation $\bar{\mathbf{X}}_{\tilde{T}} \cdot \frac{\partial \mathbf{X}_{\tilde{T}}}{\partial \boldsymbol{\psi}}$, then differentiate it again with respect to $\boldsymbol{\psi}$ getting a matrix $\bar{\bar{\boldsymbol{\psi}}}$.
6. The final estimator is $\bar{\bar{\boldsymbol{\psi}}} + \left(\frac{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})}{\partial \boldsymbol{\psi}} \right)^{\top} \left(\sum_{i=0}^h \mathbf{D}_i \right) \left(\frac{\partial (\mathbf{X}_{\tilde{T}}, \boldsymbol{\theta})}{\partial \boldsymbol{\psi}} \right)$.

As for complexity, the multiplier on the cost of simulation comes from step 5, with a factor of order $O(N)$, plus step 4, with a factor either of order $O(h \cdot N_X)$ if computed by adjoint differentiation, or of order $O(N)$ if computed by forward differentiation. The cost of steps 1 through 3 has to do with the number of discontinuities h , but is also proportional to the computational time of $\tilde{f}_i$ and $\tilde{g}$, which is usually much less relevant than the time required to perform the Euler simulation.

4.4.3 Functional Gamma

The idea of this section is that if $N_X \leq N_W$, the Feynman-Kac theorem can supply enough equations to determine the Gamma matrix $\boldsymbol{\Gamma} = \partial^2 P / \partial \mathbf{X}_0^2$ given the time decay of suitable first order sensitivities, whose computation is fast ($O(T)$) thanks to adjoint differentiation. This fact is essentially a consequence of the multi-dimensional version of a known relationship between Gamma and “functional Vega”, which is exploited for instance in Reghai et al. (2015) in a one-dimensional setting for purposes not related to the computation of a high number of sensitivities.

Indeed, it is well known that the price P , as a function of $\mathbf{X}_0$ and of the evaluation time, satisfies the PDE

$$\frac{\partial P}{\partial t} + \frac{\partial P}{\partial x} \cdot \boldsymbol{\mu} + \frac{1}{2} \text{tr} \left[\frac{\partial^2 P}{\partial x^2} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{\top} \right] - Pr + \sum C_i \delta(t - t_i) = 0,$$

where r is the risk-free rate and C_i is the cashflow paid at time t_i . Now one can note that if P' is the price for a different diffusion coefficient function Σ' , then $P' - P$ satisfies the same equation except that the discrete last addend is substituted by a continuous source term

$$\frac{1}{2} \text{tr} \left[\frac{\partial^2 P}{\partial x^2} (\Sigma' \Sigma'^\top - \Sigma \Sigma^\top) \right],$$

so that by Feynman-Kac we can conclude that

$$P'(x, t) = P(x, t) + \frac{1}{2} \mathbb{E} \left\{ \int_t^T e^{-\int_t^s r(\mathbf{X}_u, u) du} \text{tr} \left[\frac{\partial^2 P}{\partial x^2} (\Sigma' \Sigma'^\top - \Sigma \Sigma^\top) \right] (\mathbf{X}_s, s) ds \right\}.$$

Now we fix a perturbation matrix $\mathbf{H}$ and time $\tau > 0$ and choose

$$\Sigma'(t, x) = \Sigma(t, x) + \mathbf{H} I_{t < \tau};$$

differentiation with respect to $\mathbf{H}$ gives

$$\frac{\partial P'(x, 0)}{\partial \mathbf{H}} = \int_0^\tau \mathbb{E} \left\{ e^{-\int_0^s r(\mathbf{X}_u, u) du} \left[\Sigma^\top \frac{\partial^2 P}{\partial x^2} \right] (\mathbf{X}_s, s) \right\} ds;$$

finally, differentiation in $\tau = 0$ gives

$$\frac{\partial^2 P'(x, 0)}{\partial \tau \partial \mathbf{H}} = \Sigma^\top (\mathbf{X}_0, 0) \Gamma, \quad (4.11)$$

which for full rank $\Sigma(\mathbf{X}_0, 0)$ allows for the determination of Γ given $\partial^2 P'(x, 0)/(\partial \tau \partial \mathbf{H})$; we expect the computation of the latter to be faster because the second order differentiation is with respect to a one dimensional variable τ .

To compute $\partial^2 P'(x, 0)/(\partial \tau \partial \mathbf{H})$, we note that for small τ , the payoff depends on $\mathbf{H}$ only through the first simulated point $\mathbf{X}_{S_1}$. This makes the following decomposition possible:

$$\frac{\partial P'}{\partial \mathbf{H}} = \mathbb{E} \left[\frac{\partial P(\mathbf{X}_{S_1}, S_1)}{\partial \mathbf{X}_{S_1}} \cdot \frac{\partial \mathbf{X}_{S_1}}{\partial \mathbf{H}} \right], \quad (4.12)$$

where we suppose that the price P at time S_1 is a smooth function of $\mathbf{X}_{S_1}$: this is true unless the payoff has a discontinuity maturing at time S_1 , in which case one can add a point to the simulation grid $\mathbf{S}$.

If we substitute the original SDE with its Euler discretization, in which drift and diffusion coefficients are constant on time intervals $[S_i, S_{i+1})$, we get:⁶

$$\begin{aligned} \mathbf{X}_{S_1} &= (\Sigma(\mathbf{X}_0, 0) + \mathbf{H}) \mathbf{W}_\tau + \Sigma(\mathbf{X}_0, 0) (\mathbf{W}_{S_1} - \mathbf{W}_\tau) = \Sigma(\mathbf{X}_0, 0) \mathbf{W}_{S_1} + \mathbf{H} \mathbf{W}_\tau \\ &= \Sigma(\mathbf{X}_0, 0) \mathbf{W}_{S_1} + \mathbf{H} \left(S_1^{-1} \tau \mathbf{W}_{S_1} + \sqrt{S_1^{-1} (S_1 - \tau)} \tau \tilde{\mathbf{Z}} \right) \end{aligned} \quad (4.13)$$

for $\tilde{\mathbf{Z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{N_W})$ independent from $\mathbf{W}_{\mathbf{S}}$, as prescribed by the Brownian bridge distribution of $\mathbf{W}_\tau$ given $\mathbf{W}_{\mathbf{S}}$ (see e.g. Karatzas and Shreve, 1991). This means that we can rewrite (4.12) as

$$\frac{\partial P'}{\partial \mathbf{H}} = \mathbb{E} [\mathbf{W}_\tau \cdot \nabla P(\mathbf{X}_{S_1}, S_1)^\top] = \mathbb{E} \left[\left(\frac{\tau}{S_1} \mathbf{W}_{S_1} + \sqrt{\frac{(S_1 - \tau)\tau}{S_1}} \tilde{\mathbf{Z}} \right) \cdot \nabla P(\mathbf{X}_{S_1}, S_1)^\top \right].$$

⁶This way we differentiate the price *as approximated by the pricing engine*; this should make little difference in practice, and is anyway arguably even more useful than the differentiation of the abstract price which would arise from the theoretical exact solution of the SDE. Note that the argument in fact needs this approximation only in the first time interval $[0, S_1)$.

This expression is not yet ready for ordinary differentiation with respect to τ because of the $\sqrt{\tau}$ dependence in the second addend. However, since we are only concerned with the value of the right hand side in $\mathbf{H} = \mathbf{0}$, in which point $\mathbf{X}_{S_1}$ is a function of $\mathbf{W}_{S_1}$ but not of $\tilde{\mathbf{Z}}$, we can use independence to deduce

$$\mathbb{E} \left[\sqrt{\frac{(S_1 - \tau)\tau}{S_1}} \tilde{\mathbf{Z}} \cdot \nabla P(\mathbf{X}_{S_1}, S_1)^\top \right] = \mathbb{E} \left[\sqrt{\frac{(S_1 - \tau)\tau}{S_1}} \tilde{\mathbf{Z}} \right] \cdot \mathbb{E} [\nabla P(\mathbf{X}_{S_1}, S_1)^\top] = \mathbf{0}.$$

To sum up, to apply (4.11), we have to estimate simply

$$\frac{\partial}{\partial \tau} \mathbb{E} \left[\frac{\tau}{S_1} \mathbf{W}_{S_1} \cdot \nabla P(\mathbf{X}_{S_1}, S_1)^\top \right] = \mathbb{E} \left[\frac{\mathbf{W}_{S_1}}{S_1} \cdot \bar{\mathbf{X}}_{S_1} \right] \quad (4.14)$$

where in the second equality we have substituted the Delta sensitivity ∇P with any estimator thereof.

The final equation (4.14) closely resembles that of likelihood-ratio based estimators like LRMAAD, while being both simpler to implement since it does not involve second order differentiation of the drift and diffusion coefficients, and more flexible since it does not prescribe which first order estimator $\bar{\mathbf{X}}_{S_1}$ to use. This also means that it shares the main drawback of likelihood ratio weights, i.e. the reciprocal dependence on the small discretization step S_1 , which may impact negatively the Monte Carlo variance. However, as is the case in vibrato, a simple antithetic trick can save the day, since the resulting estimator becomes

$$\Sigma^\top(\mathbf{X}_0, 0)^{-1} \frac{\mathbf{W}_{S_1}}{2S_1} \cdot [\bar{\mathbf{X}}_{S_1}(\mathbf{W}_{S_1}) - \bar{\mathbf{X}}_{S_1}(-\mathbf{W}_{S_1})], \quad (4.15)$$

where, roughly speaking, we expect

$$\mathbf{W}_{S_1} = O(\sqrt{S_1}) \quad \text{and} \quad \bar{\mathbf{X}}_{S_1}(\mathbf{W}_{S_1}) - \bar{\mathbf{X}}_{S_1}(-\mathbf{W}_{S_1}) = O(\mathbf{W}_{S_1}) = O(\sqrt{S_1}),$$

jointly counterbalancing the S_1^{-1} factor.

We stress that for maximum effectiveness of this variant, one should change the sign only of $\mathbf{W}_{S_1}$, unlike in standard antithetic Monte Carlo which inverts the full Brownian path. This does not prevent the usage of ordinary antithetic trajectories on top of the final estimator (4.15), as one would do for any other method.

4.5 Empirical results

In this section, we will compare empirically the above algorithms on several test cases. We will always suppose that the payoff is written on one or more underlying assets A_i with the simplest Black-Scholes dynamics

$$A_i = \exp(X_i), \quad dX_i = -\frac{\sigma_i^2}{2} dt + \sigma_i dB_i, \quad d\langle B_i, B_j \rangle = \rho_{ij} dt, \quad (4.16)$$

where B_i are correlated Brownian motions; of course, in this stylized case Euler simulation is not strictly necessary since finite-step transition probabilities are available in closed form, but we neglect this simplification which would not apply to general diffusive models. Note that (4.16) is readily expressed in the form (4.1), which formally prescribes independent Brownian motions, by a Choleski factorization of the

instantaneous correlation matrix $\mathbf{R} = (\rho_{ij})_{i,j \leq N_W}$:

$$\begin{aligned} d\mathbf{X} &= \boldsymbol{\mu} dt + \boldsymbol{\Sigma} d\mathbf{W}, \\ \boldsymbol{\mu} &:= (-\sigma_i^2/2)_{i \leq N_W} \quad \text{and} \quad \boldsymbol{\Sigma} := \text{diag}(\sigma_i)_{i \leq N_W} \cdot \text{Choleski}(\mathbf{R}). \end{aligned}$$

The following names will be used to denote subsets of the full Hessian $\mathbf{H}$:

$$\text{Gamma} = \left(\frac{\partial^2 P}{\partial X_i(0) \partial X_j(0)} \right)_{i,j}, \quad \text{Vanna} = \left(\frac{\partial^2 P}{\partial X_i(0) \partial \sigma_j} \right)_{i,j}, \quad \text{Volga} = \left(\frac{\partial^2 P}{\partial \sigma_i \partial \sigma_j} \right)_{i,j};$$

this is slight abuse of terminology, since to adhere to common jargon one should substitute in the definitions $\mathbf{X}$ with $\mathbf{A}$, but in our numerical analysis we prefer to display the sensitivity with respect to the simulated driver $\mathbf{X}$, because for most estimators this is more natural. One can always obtain the sensitivity to $\mathbf{A}$ plugging the result into the elementary relations

$$\begin{aligned} \frac{\partial^2 P}{\partial A_i(0) \partial A_j(0)} &= \frac{1}{A_i(0) A_j(0)} \frac{\partial^2 P}{\partial X_i(0) \partial X_j(0)} - \delta_{ij} \frac{1}{A_i(0)^2} \frac{\partial P}{\partial X_i(0)}, \\ \frac{\partial^2 P}{\partial A_i(0) \partial \sigma_j} &= \frac{1}{A_i(0)} \frac{\partial^2 P}{\partial X_i(0) \partial \sigma_j}. \end{aligned}$$

In the sequel, all algorithms will be denoted by acronyms: we point to Tables 4.1 and 4.2 for easy reference. Note that FT can be based on any first order estimator, hence we will use the notations FT-PW, FT-VB, FT-OPP, FT-DAAD to denote the functional Gamma method of Section 4.4.3 where the first order estimator used to compute $\bar{\mathbf{X}}_{S_1}$ is respectively given by pathwise differentiation, vibrato Monte Carlo, Optimal Partial Proxy and DAAD.

All the test payouts will be autonomous in the sense of (4.4), so that all methods described above apply; only when the payoff is discontinuous, some of them will be unavoidably excluded. In fact in those cases one might smooth the payoff, not only to enable otherwise unusable estimators, but also to try and improve the performance of other ones. We neglect this possibility since smoothing techniques would involve a careful choice of their parameters to correctly balance the variance gain with the bias introduced, which is problematic already for first order estimators: see the analysis in Section 3.6.

In fact, also the choice of the displacement for finite difference based algorithms (FDIFF2, FDPW, VFD, FDDAAD) may be a serious concern. We will use for $\Delta X_i(0)$ and $\Delta \sigma_i$ the largest power of 10 such that the confidence interval of the finite difference estimator contains the analytical result when available, or does not appear to diverge significantly from the confidence intervals of unbiased methods otherwise. This is a bit of cheating, and is only justified because we are just looking for benchmarks to the analytical methods which are the main subject of our analysis; however, if one would actually choose a finite difference method as his only sensitivity calculation engine, such comparison terms would not be available, and much attention should be paid to accurate heuristics for the choice of a good finite shift.

4.5.1 Single-asset payoffs: stability of the estimators

Before analysing settings in which the high number of sensitivities to compute is a major driver in the choice of the algorithm, we deem it useful to explore the behaviour of the several estimators in the simpler task of computing the Gamma sensitivity

Method	$T = 1d$	$T = 1y$
FDIFF2	48.562839	3.604733
FDPW	41.621283	3.097918
HOPP	47.695944	3.719352
LRMAAD	104.264491	12.588643
AADLRM	104.264491	12.588643
VAD	23.330228	1.881182
DAAD2	19.015445	1.552514
FT-PW	53.021142	6.652639
FT-VB	65.882049	8.471453

Table 4.3: Gamma sensitivity of a vanilla call: standard deviation for 10^5 Monte Carlo paths, averaged over 40 values of A_0 .

with respect to a single underlying asset. This way, we can for a moment forget about the computational burden, which is comparable and affordable for all methods, and concentrate on more intrinsic and implementation-independent properties, among which we put the focus on the following:

- Variance: how large is the confidence interval for a fixed number of Monte Carlo paths? We will analyse the standard deviation of the result for 10^5 simulated paths, which by the central limit theorem should be proportional to the width of a centred confidence interval for any confidence level desired.
- Stability: does the result change smoothly while moving the initial condition S_0 ? This is a highly desirable property in practice, since one does not want the Greeks to jump wildly as the market evolves.

In all experiments, the volatility parameter is set to $\sigma = 0.2$ in yearly time units.

Call option

As a prototypical example of a continuous payoff we take a vanilla call option $(A_T - K)^+$ with strike $K = 100$, and vary the initial condition A_0 and the maturity time T . In particular, we take both a moderate maturity T of one year, divided for the purpose of Euler simulation into 100 equal time intervals; and a short maturity $T = 1$ day, with hourly steps in the Euler simulation. The analytical price is well known and will serve as a check.

First of all, we compare the Monte Carlo uncertainties in Table 4.3. We make two main remarks:

- Likelihood ratio based methods have by far the largest confidence intervals, followed by functional Gamma.
- For this simple payoff, only DAAD2 and VAD offer a material improvement over naive finite differences.

As far as smooth dependence on A_0 is concerned, we first consider the case in which T is one year. Figure 4.1 plots the complete results: all methods appear to work reasonably well, but those based on likelihood ratio or functional Gamma display larger errors, because of the larger confidence intervals. The smoothest analytical methods

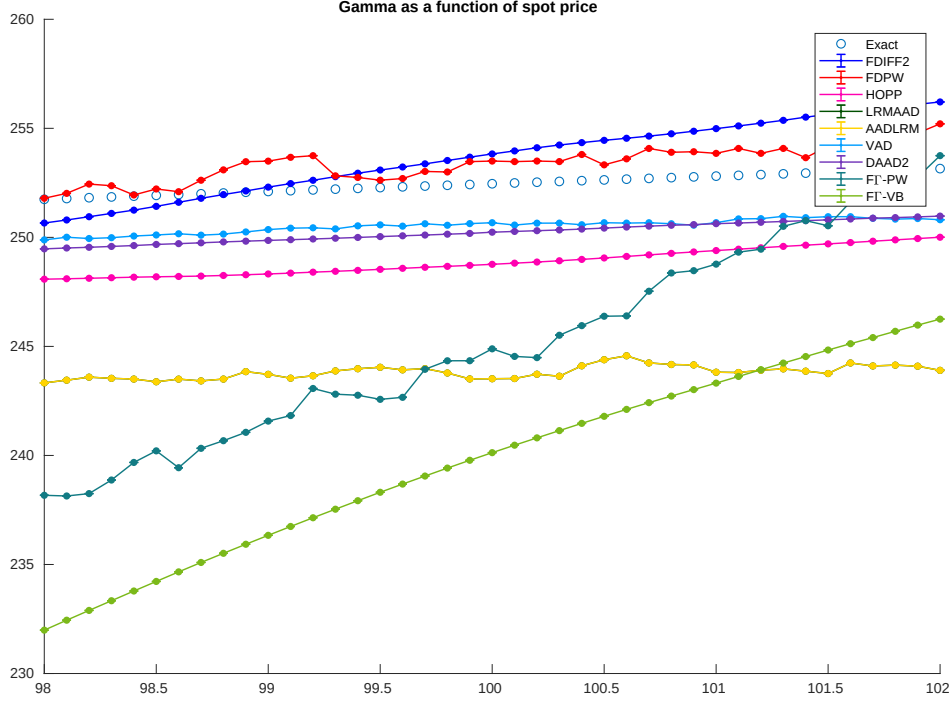


Figure 4.1: Gamma sensitivity of a vanilla call with maturity $T = 1y$ and strike $K = 100$ as a function of the spot price A_0 : all estimators, 10^5 paths.

are HOPP and DAAD2; also VAD performs well, while FDPW is quite more erratic. However, one should again note that for this simple payoff, elementary second order finite differences FDIFF2, here with displacement $\Delta X_0 = 0.01$, are not that bad.

We tried to stress the exercise considering an option expiring in $T = 1$ day: the results are in Figure 4.2. One notable difference is that FDIFF2 ($\Delta X_0 = 0.001$) and HOPP lose some smoothness, while DAAD2 still follows the shape of the exact solution remarkably well. Moreover, FT-VB becomes the most stable alternative, even though it is still significantly more noisy than most of its competitors, as we saw in Table 4.3. This smoothness is not observed in FT-PW, showing that the flexibility in the choice of the underlying first order estimator which functional Gamma offers over traditional likelihood ratio methods is a valuable asset.

Digital option

The archetypal example of a discontinuous payoff is the digital option $I_{A_T > K}$, for which we repeat the analyses of the previous subsection. Note that several methods cannot be used any more.

The detail of Monte Carlo uncertainties is in Table 4.4. The winners are FT-DAAD, DAAD2 and FT-VB, followed by the biased VFD; all methods are much better than FDIFF2, so we exclude it from the graphs of this subsection.

Figure 4.3 shows that for $T = 1$ year, DAAD based methods (DAAD2, FDDAAD, FT-DAAD) display the smoothest dependence on A_0 . The same remarks apply to $T = 1$ day, which was plotted in Figure 4.4, where the analytical exact Gamma has been subtracted from the estimated value for a better display cleaned from the trend.

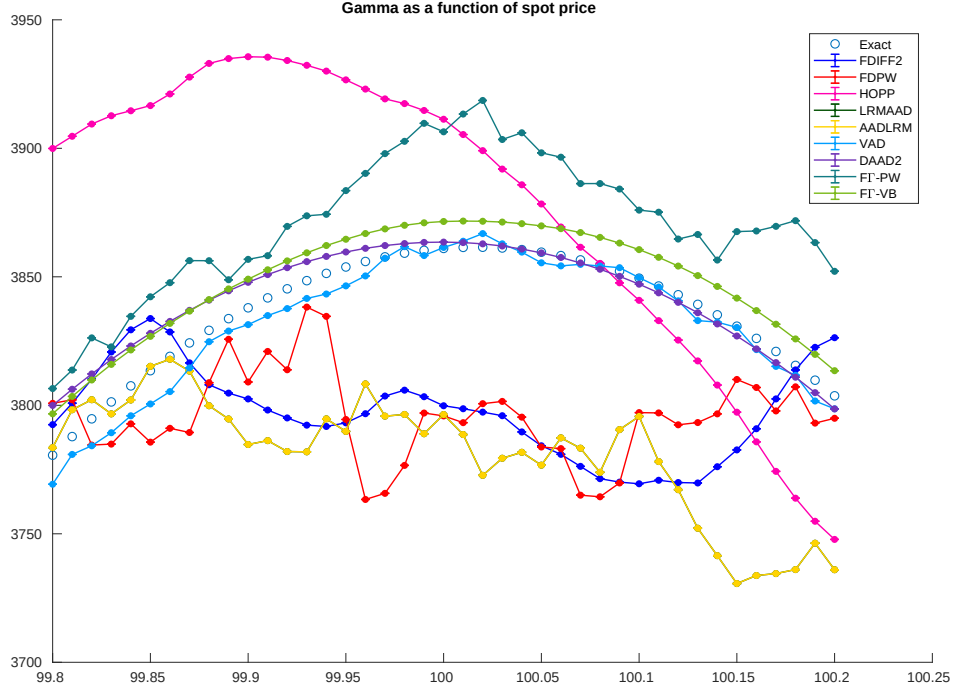


Figure 4.2: Gamma sensitivity of a vanilla call with maturity $T = 1d$ and strike $K = 100$ as a function of the spot price A_0 : all estimators, 10^5 paths.

Method	$T = 1d$	$T = 1y$
FDIFF2	865.468515	6.303787
HOPP	198.053629	1.672700
VFD	88.642863	0.699658
FDDAAD	134.068248	1.095060
DAAD2	73.659168	0.587246
FT-VB	75.564853	0.652947
FT-OPP	123.806197	1.030655
FT-DAAD	61.683419	0.547219

Table 4.4: Gamma sensitivity of a digital call: standard deviation for 10^5 Monte Carlo paths, averaged over 40 values of A_0 .

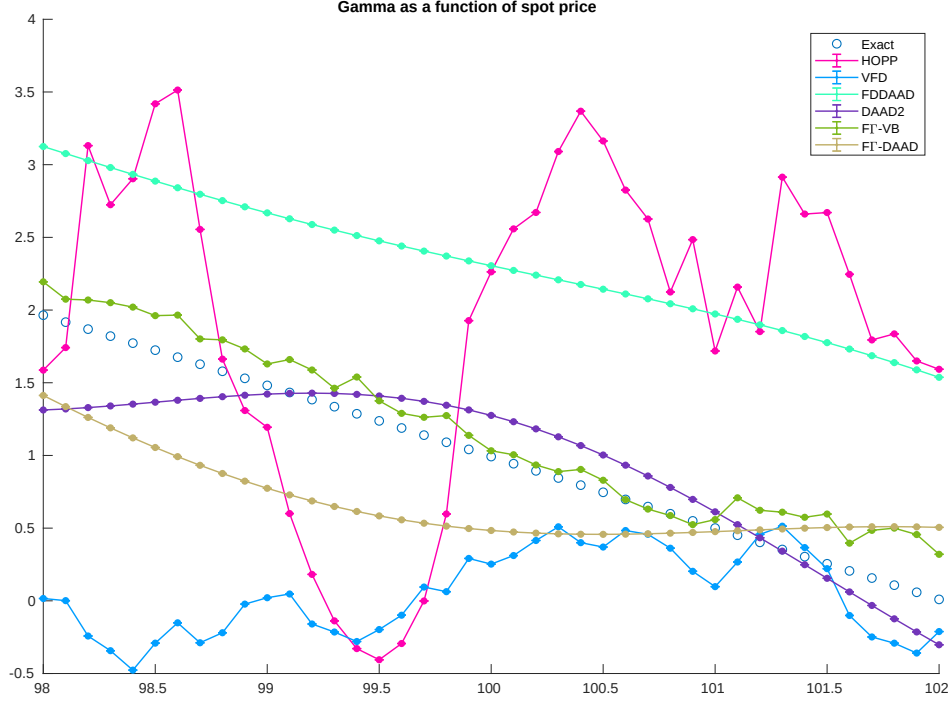


Figure 4.3: Gamma sensitivity of a digital call with maturity $T = 1y$ and strike $K = 100$ as a function of the spot price A_0 : low-variance estimators, 10^5 paths.

4.5.2 Multi-asset payoffs: efficiency of the algorithms

We now move to multi-asset payoffs, where the full Hessian $\mathbf{H}$ has a high number of entries. This imposes a rethinking of how the various estimators are compared, because a more noisy one may still be preferable if it is significantly faster. We believe that the most objective metric is therefore the estimated run time which would be needed to achieve a unit Monte Carlo variance. This is easily computed as

$$(\text{run time}) \times (\text{estimated standard deviation})^2, \quad (4.17)$$

since it is well known that to reduce the uncertainty by a factor of two one should increase the number of simulated paths by a factor of four.

In all examples, all assets have spot value $A_i(0) = 100$ and volatility parameter $\sigma_i = 0.2$, while the instantaneous correlation between couples of Brownian drivers is set to a flat value $\rho_{ij} = 0.5$. Each method was tested on a Monte Carlo simulation with 10^4 paths.

We will plot the chosen metric (4.17) separately for every second order sensitivity in the Hessian matrix, because while on the one hand the run time is a common number, on the other hand the standard deviation is different entry by entry. In all the figures, the graphs on the left concern the sensitivities in which both differentiation variables refer to the same asset, while all other entries are plotted in the graphs on the right.



Figure 4.4: Gamma sensitivity of a digital call with maturity $T = 1d$ and strike $K = 100$ as a function of the spot price A_0 : residual of low-variance estimators with respect to the analytical value, 10^5 paths.

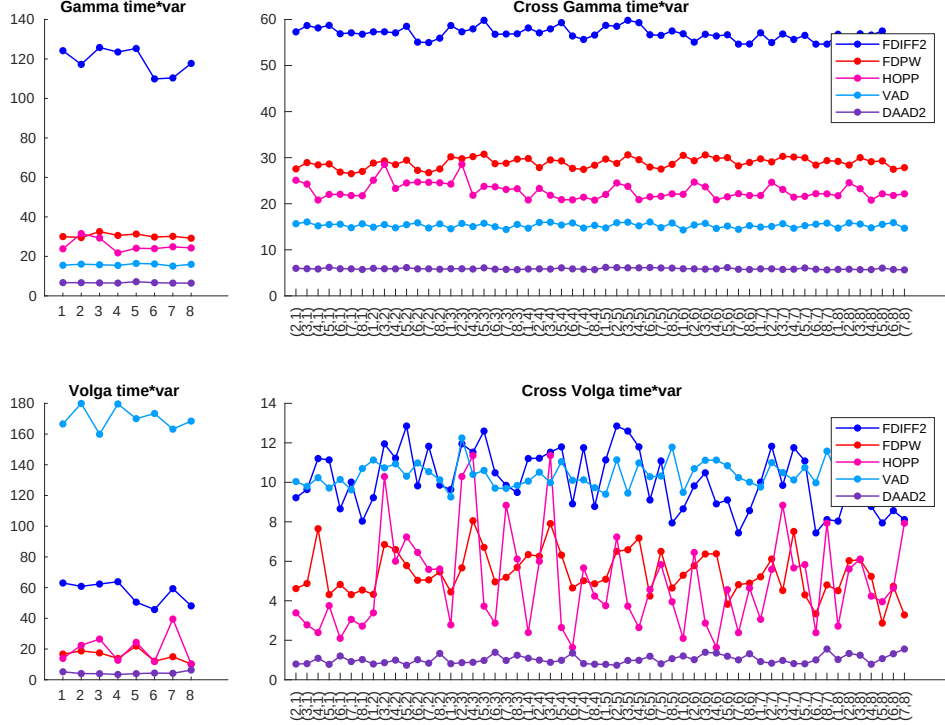


Figure 4.5: Gamma and Volga sensitivities of a call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of all methods except likelihood ratio (out of scale).

Basket call

The multidimensional generalization of the first example of Section 4.5.1 is a call on the average of several assets:

$$\left(\frac{1}{N_X} \sum_{i=1}^{N_X} A_i(T) - K \right)^+,$$

which we will price for $N_X = 8$, $T = 1$ year and $K = 100$.

The results are in Figures 4.5 and 4.6; note the absence of likelihood ratio methods, since their values are between one and two orders of magnitude larger than those of the other algorithms, and the graph would be unreadable. In particular, finite differences should be preferred to likelihood ratio regardless of their higher cost per path; while all other methods outperform FDIFF2, as the plot shows. DAAD2 is consistently much more effective than its competitors, while the second-best depends on the sensitivity of interest: it is VAD for Gamma, but becomes FDPW for Volga, VAD for Delta of Vega, HOPP for Vega of Delta (although Delta of Vega and Vega of Delta estimate theoretically the same number by the Schwarz theorem, see the last paragraph of this section).

If one wants to compute only Gamma, then functional methods are also available; moreover, LRMAAD might in principle become competitive thanks to its efficient implementation of Section 4.3.2. For a fair comparison we updated also the run times

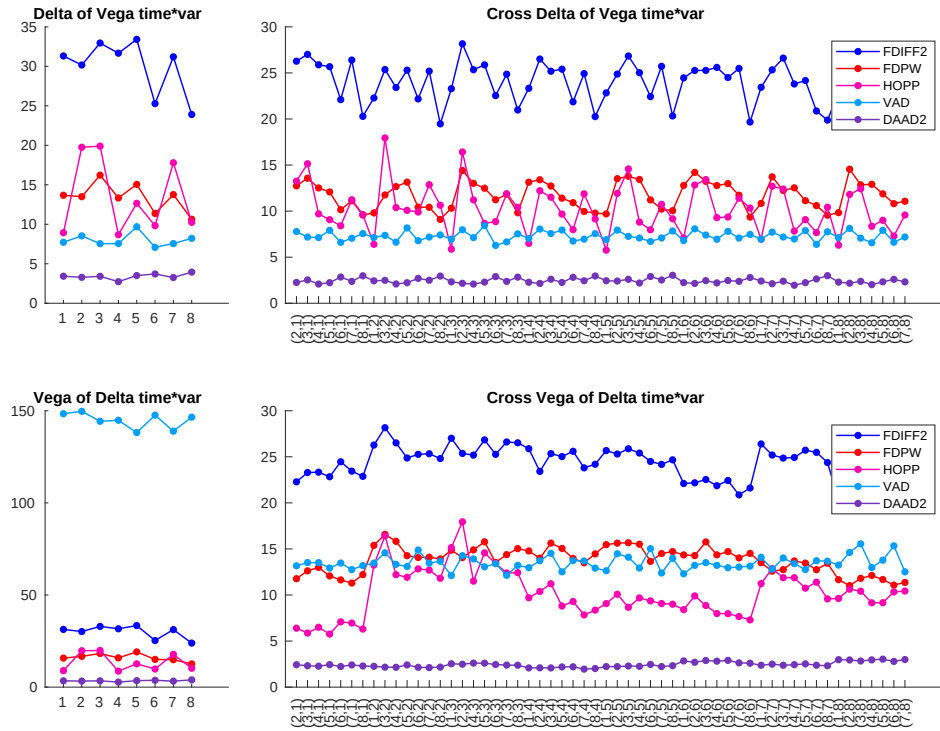


Figure 4.6: Vanna sensitivities of a call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of all methods except likelihood ratio (out of scale).

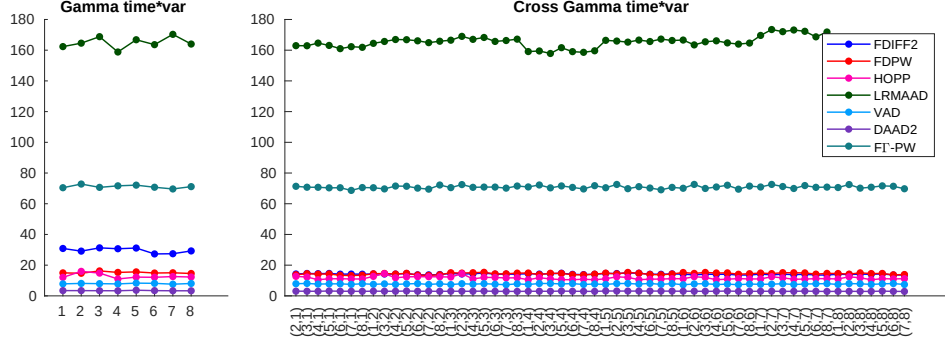


Figure 4.7: Gamma sensitivities of a call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of all methods except AADLRM (out of scale).

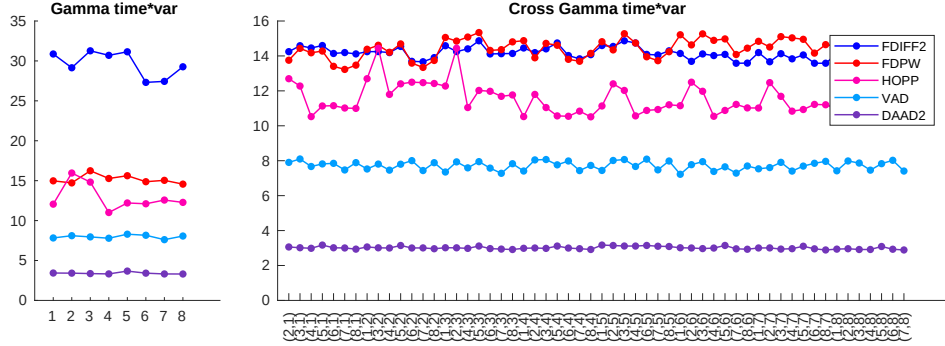


Figure 4.8: Gamma sensitivities of a call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of all methods except likelihood ratio and functional Gamma (out of scale).

of the other methods, which are smaller if sensitivities to σ_i are not required. Our empirical findings in this setting are in Figure 4.7, and are not encouraging: LRMAAD is still far too noisy even considering that it is much faster than linear-cost methods (not to speak of FDIFF2), and FI, while significantly better, does not improve the variance sufficiently. We remark that now that the number of required sensitivities is “only” 64, beating the dumb FDIFF2 benchmark is harder than one might expect, and off the diagonal only DAAD2 is a neat improvement (Figure 4.8, right-hand graph).

Finally, before accepting this disappointing remark, we note that most of the analytical estimators we test produce asymmetric matrices, although we know that Hessians are symmetric: so we have two distinct estimators for each off-diagonal entry. Therefore, we can look for a linear combination to reduce the Monte Carlo variance: equivalently, we can correct the estimator Γ_{ij} using the mean-zero $(\Gamma_{ji} - \Gamma_{ij})$ as a control variate. The coefficients will be chosen by a small-sample preliminary Monte Carlo run as is customary for control variates (Glasserman, 2004). Some methods benefit from this trick more than others, as one can see from Figure 4.9: in particular, functional Gamma comes closer to finite differences (but is still more than twice slower in getting unit variance), and VAD becomes as good as DAAD2 on Cross Gammas.

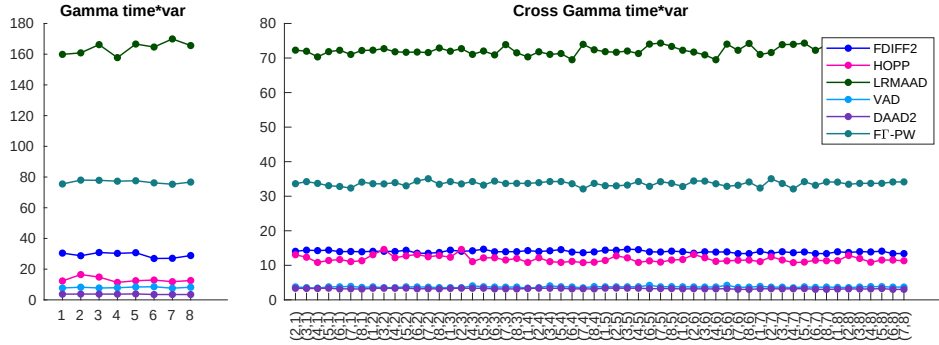


Figure 4.9: Gamma sensitivities of a call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of FDIFF2 and of symmetrized analytical estimators except AADLRM (out of scale).

Basket digital

A good representative example of a discontinuous multi-asset product is a digital option on a basket:

$$I_{(K, +\infty)} \left(\frac{1}{N_X} \sum_{i=1}^{N_X} A_i(T) \right).$$

As in the previous subsection, we take $N_X = 8$, $T = 1$ year and $K = 100$.

The results for the full Hessian are in Figures 4.10 and 4.11: they show that FDIFF2 is inadequate for discontinuous multi-asset payoffs, while FDDAAD and HOPP are not optimal either. DAAD2 and VFD are comparable.

The biggest surprise, however, comes from the estimators restricted to the Gamma matrix in Figure 4.12. Indeed, we see that the best linear-cost methods DAAD2 and VFD are clearly overcome by constant-cost functional Gamma methods, with FI'-DAAD being 4-5 times faster than the fastest linear-complexity algorithm (always in uncertainty-adjusted terms). This is at odds with the results obtained in the previous subsection on the “easy” call payoff, probably because on continuous payouts traditional methods already work very well.

4.6 Conclusions

In this chapter, we have compared a wide range of old and new methods for the computation of second order sensitivities of Monte Carlo prices of financial products, with special emphasis on algorithmic efficiency when a large number of Greeks must be computed.

From the literature review viewpoint, the theoretical analysis has examined, with extensions (LRMAAD, VAD) and variations (VFD), the main algorithms to compute the full Hessian matrix of an expected value when the underlying is driven by a multidimensional Brownian diffusion. Several limitations of applicability are found:

1. Many methods often have too high Monte Carlo uncertainties: in particular, those based on likelihood ratio (LRMAAD, AADLRM), and to a lesser extent those based on finite differences of the pathwise estimator (FDPW).

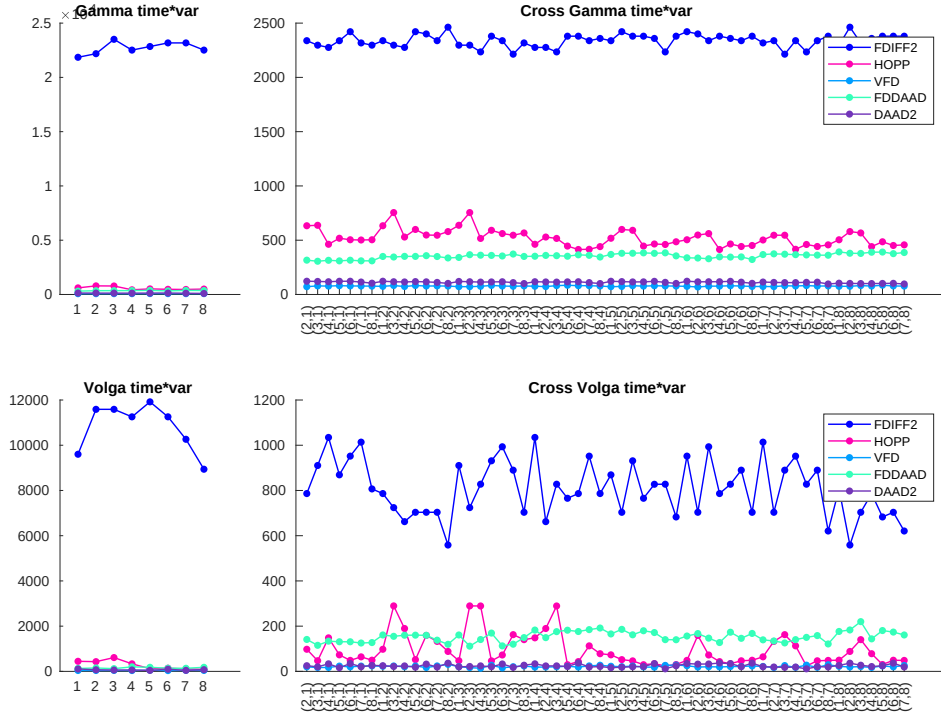


Figure 4.10: Gamma and Volga sensitivities of a digital call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of all methods.

2. Most methods only apply when there is at least one random driver per underlying ($N_X \leq N_W$) and the payoff does not depend directly on the initial state $\mathbf{X}_0$, nor on the model parameters $\boldsymbol{\theta}$ (the exceptions being HOPP and the dumb FDPW).
3. Most methods cannot handle discontinuous payoffs (the exceptions being HOPP and VFD).
4. For all methods except LRMAAD restricted to sensitivities to $\mathbf{X}_0$ only, the multiplicative overhead on the run time for only-price grows linearly in the number of rows of the Hessian.

The original estimators of this chapter should be considered in view of the above limitations of their many competitors (see Table 4.5 for a visual summary):

- DAAD2 is the only method besides HOPP which solves simultaneously issues 1, 2 and 3, and in our tests was always significantly more effective than the latter from the statistical uncertainty point of view, while showing remarkable smoothness properties with respect to small changes in the input parameters.
- FT is the first estimator solving simultaneously issues 1 and 4 in at least one relevant case; it is also the first constant-complexity algorithm applicable to discontinuous payoffs. Its best-of-class efficacy in the basket digital test case, coupled with the easy implementation, qualify it as a must-try when tackling a new tough high-dimensional problem. Since it is in fact a meta-algorithm based on any first order

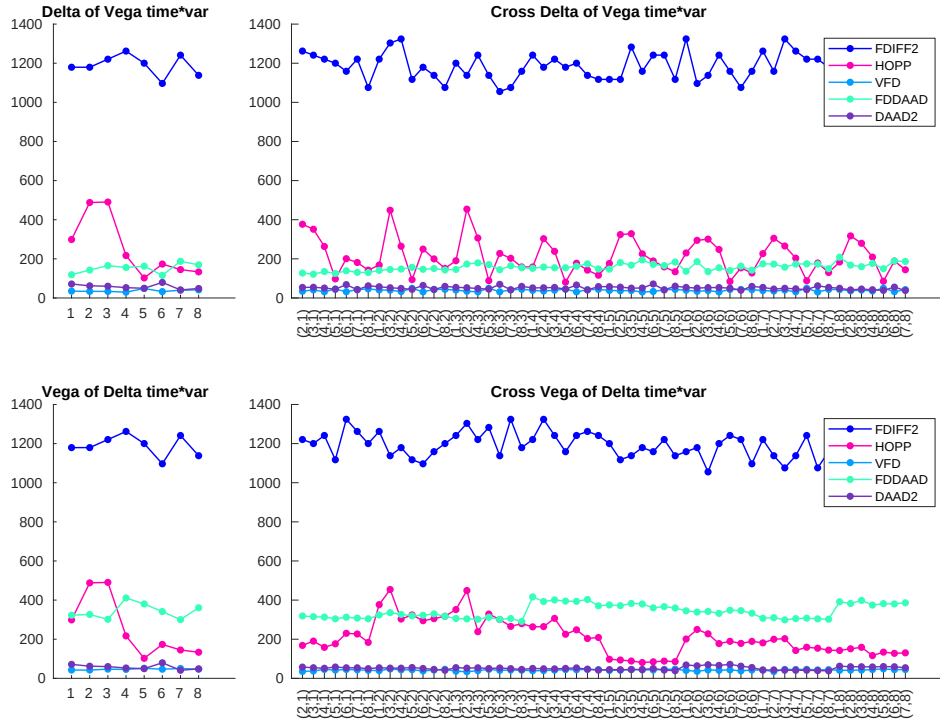


Figure 4.11: Vanna sensitivities of a digital call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of all methods.

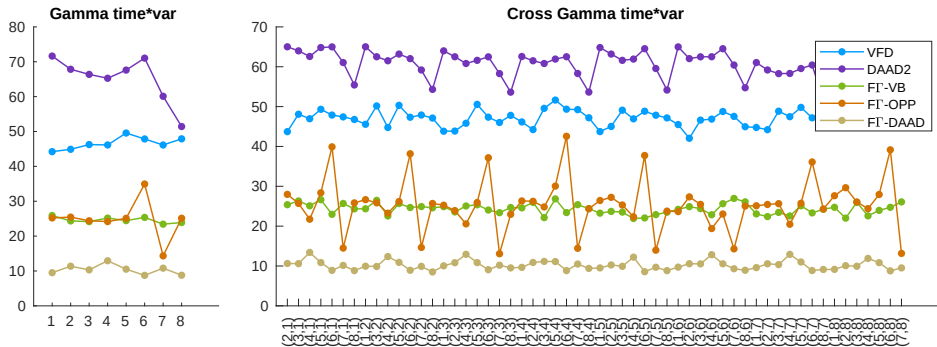


Figure 4.12: Gamma sensitivities of a digital call with maturity $T = 1y$ and strike $K = 100$ on a basket of 8 underlyings: efficiency of best-performing methods.

Table 4.5: Properties and domain of applicability of the methods.

	FDPW	LRMAAD	AADLRM	VAD	HOPP	DAAD2	FT
Low variance	✗	✗	✗	✓	✓	✓	✗/✓ ¹
Non-autonomous	✓	✗	✗	✗	✓	✓	✓
Discontinuous	✗	✗	✗	✓ ²	✓	✓	✓ ³
Constant-cost Γ	✗	✓	✗	✗	✗	✗	✓
Full Hessian	✓	✓	✓	✓	✓	✓	✗

¹ Depending on the payoff.² In fact one must substitute AD with a finite difference (VFD).³ If handled by the underlying algorithm.

estimator, it offers significant flexibility for either seamless introduction in any financial library already capable of computing first order sensitivities, or for fine tuning to specific payoffs.

Besides systematic testing on a large range of payoff types and model dynamics, a couple of issues remain open for further research. In the field of linear-cost methods, the main challenge is related to payouts with many discontinuities, for which the computational complexity of the adjustments becomes relevant, particularly when it scales quadratically in the number of digital features as for the current implementation of DAAD2. Moreover, this chapter opens the new field of constant-cost methods, where a general-purpose algorithm is still to be found, since FT in one example showed poor performance, and is anyway limited to derivatives with respect to $\mathbf{X}_0$. Finally, one might consider the possibly varied effectiveness of variance reduction techniques to different estimators.

Appendix A

Technical appendix

A.1 Hedging with any number of correlated assets

In this section we generalize the formulas of Chapter 1 to the case when more than two risk factors and hedging instruments are involved. We will follow the local variance minimization argument, but the global one can be reproduced as well.

Only for this section, the derivative price Π_t will be a function of n asset prices:

$$\Pi_t = \Pi \left(S_t^{(1)}, \dots, S_t^{(n)} \right),$$

where the assets have dynamics

$$dS_t^{(i)} = \mu^{(i)}(t) dt + \sigma^{(i)}(t) dW_t^{(i)};$$

we assume that $\mathbf{W}_t = (W_t^{(1)}, \dots, W_t^{(n)})$ is a vector-valued correlated Wiener process with instantaneous correlation matrix $\boldsymbol{\rho}$:

$$\mathbb{E} \left[dW_t^{(i)} dW_t^{(j)} \right] = \rho_{ij} dt.$$

We also define the instantaneous covariance matrix $\mathbf{C}$ by

$$C_{ij}(t) dt = E \left[dS_t^{(i)} dS_t^{(j)} \right], \quad \text{i.e.} \quad C_{ij}(t) = \sigma^{(i)}(t) \sigma^{(j)}(t) \rho_{ij}.$$

For a generic (column) vector quantity $\mathbf{X} = (X^{(i)})_{i \in \{1, \dots, n\}}$, and a generic $A \subseteq \{1, \dots, n\}$, we adopt the notation $\mathbf{X}^A$ to denote the vector whose components are $X^{(i)}$ for $i \in A$; analogously, for any matrix $\mathbf{M}$, $\mathbf{M}_{A,B}$ will denote the submatrix obtained considering only rows whose index is in A and columns whose index is in B .

Let us assume that, at some time t , and for each $i = 1, \dots, n$, an amount $\Delta^{(i)}$ of the i -th asset is held in the hedging portfolio; let us further suppose that we consider, among all the possible actions on the portfolio, only the ones involving some subset of the assets. The set of assets eligible for rebalancing will be represented by the set of their indices $A \subseteq \{1, \dots, n\}$ ($A = \emptyset$ representing the choice not to rehedged at all, and $A = \{1, \dots, n\}$ representing complete freedom in the rehedging choice). We further use the letter B to denote the complementary of A : $B = \{1, \dots, n\} \setminus A$.

We then want to choose a new value of $\boldsymbol{\Delta}^A$ in such a way to minimize

$$\text{Var} [d\Pi_t - \boldsymbol{\Delta} \cdot d\mathbf{S}_t];$$

with the notation $\mathbf{a} := (\nabla \Pi_t - \Delta)^A$, $\mathbf{b} := (\nabla \Pi_t - \Delta)^B$, this variance reduces to

$$(\mathbf{a}^\top \mathbf{C}_{A,A} \mathbf{a} + 2\mathbf{b}^\top \mathbf{C}_{B,A} \mathbf{a} + \mathbf{b}^\top \mathbf{C}_{B,B} \mathbf{b}) dt, \quad (\text{A.1})$$

to be minimized in the free variable $\mathbf{a}$. It is not hard to see that the solution is

$$\mathbf{a} = -\mathbf{C}_{A,A}^{-1} \mathbf{C}_{A,B} \mathbf{b}, \quad \text{i.e.} \quad \Delta^A = \nabla^A \Pi_t + \mathbf{C}_{A,A}^{-1} \mathbf{C}_{A,B} (\nabla^B \Pi_t - \Delta^B), \quad (\text{A.2})$$

which substituted into (A.1) gives a residual variance (for a small elapsed time δ)

$$V_{t+}(A) = [\mathbf{b}^\top (\mathbf{C}_{B,B} - \mathbf{C}_{B,A}(\mathbf{C}_{A,A})^{-1} \mathbf{C}_{A,B}) \mathbf{b}] \delta,$$

and hence a variance reduction

$$\Delta_A V = V_{t-} - V_{t+}(A) = [\mathbf{a}^\top \mathbf{C}_{A,A} \mathbf{a} + 2\mathbf{b}^\top \mathbf{C}_{B,A} \mathbf{a} + \mathbf{b}^\top \mathbf{C}_{B,A} (\mathbf{C}_{A,A})^{-1} \mathbf{C}_{A,B} \mathbf{b}]_{t-} \delta,$$

differing from (A.1) only in that $\mathbf{C}_{B,B}$ has been substituted by $\mathbf{C}_{B,A}(\mathbf{C}_{A,A})^{-1} \mathbf{C}_{A,B}$.

If buying or selling the i -th asset incurs a cost c_i , then reheding all assets in A will give rise to a total cost $c_A = \sum_{i \in A} c_i$. If we keep the utility function U satisfying (1.4), then after reheding on subset A we have

$$U_{t+}(A) = -\exp \left\{ \frac{\lambda^2}{2} (V - \Delta_A V) - \lambda (M - c_A) \right\} = U_{t-} \cdot \exp \left\{ \lambda \left(c_A - \frac{\lambda}{2} \Delta_A V \right) \right\}.$$

Again, the attractiveness of the A “move” is higher if the exponential multiplying the negative U_{t-} is lower. We end up with a measure of the attractiveness of A -reheding defined by

$$G_A = \frac{\lambda}{2} \Delta_A V - c_A = \frac{\lambda \delta}{2} [\mathbf{a}^\top \mathbf{C}_{A,A} \mathbf{a} + 2\mathbf{b}^\top \mathbf{C}_{B,A} \mathbf{a} + \mathbf{b}^\top \mathbf{C}_{B,A} (\mathbf{C}_{A,A})^{-1} \mathbf{C}_{A,B} \mathbf{b} - \Theta_A], \quad (\text{A.3})$$

where $\Theta_A := \sum_{i \in A} \Theta_i$, with

$$\Theta_i := \frac{2}{\lambda \delta} c_i$$

equal to the quantities defined in (1.6) which describe the behaviour of the trader in the single asset case.

To sum up, at each time and for each $A \subseteq \{1, \dots, n\}$ ($A = \emptyset$ included), the trader is able to compute $G_A(\lambda \delta)^{-1}$ from (A.3) in terms of known quantities; he will choose the subset with highest such score, and if it is non empty, he will adjust his portfolio according to (A.2).

A.2 First order distributional differentiation

The purpose of this appendix is to provide rigorous hypotheses and proofs for the main formulas whose derivation was heuristic in the Chapter 3. Unless otherwise stated, all random variables and L^p spaces are with respect to a fixed probability measure.

Let us begin with a lemma covering the unidimensional case. Given $p \in [1, \infty]$, we will denote by p' its conjugate exponent, defined by the equation $1/p + 1/p' = 1$.

Lemma A.1. *Let X be a real-valued random variable with continuous density ρ , and let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a function such that $g(X) \in L^1$. Fix θ_0 in $\mathbb{R}$, and let $f = f(x, \theta)$*

be a continuous function such that $f(\cdot, \theta_0)$ is zero only in a finite number of points $x_1 < \dots < x_k$; f is C^1 in a neighbourhood of (x_i, θ_0) for each i ; and

$$\frac{\partial f}{\partial x}(x_i, \theta_0) \neq 0 \quad \text{and} \quad g \text{ is continuous in a neighbourhood of } x_i.$$

Suppose moreover that at least one of the following hypotheses is satisfied:

1. $f(\cdot, \theta)$ has no more than k zeros for each θ in a neighbourhood of θ_0 , or:
2. $\liminf_{x \rightarrow \infty} |f(x, \theta_0)| > 0$ and there exists $p > 1$ such that the following integrability conditions hold:

- $g(X) \in L^p$;
- $|f(x, \theta_0 + \Delta\theta) - f(x, \theta_0)| \leq L(x)|\Delta\theta|$ for each $|\Delta\theta| < \delta$, where δ is a positive constant and L is a function such that $L(X) \in L^{p'}$.

Then

$$\frac{\partial}{\partial \theta} \bigg|_{\theta=\theta_0} \mathbb{E}[g(X)I_{f(X,\theta)>0}] = \sum_{i=1}^k \left| \frac{\partial f}{\partial x}(x_i, \theta_0) \right|^{-1} \frac{\partial f}{\partial \theta}(x_i, \theta_0) g(x_i) \rho(x_i). \quad (\text{A.4})$$

Proof. Let $\varepsilon > 0$ be arbitrary, and define $m = \liminf_{x \rightarrow \infty} |f(x, \theta_0)|$.

By the implicit function theorem, there exist $\delta' < \delta$ and disjoint open intervals $U_1, \dots, U_k$ such that

$$\{f = 0\} \cap (U_i \times (\theta_0 - \delta', \theta_0 + \delta')) = \{(\xi_i(\theta), \theta) : \theta \in (\theta_0 - \delta', \theta_0 + \delta')\}, \quad (\text{A.5})$$

where $\xi_i(\theta_0) = x_i$ and

$$\xi'_i(\theta_0) = - \left[\frac{\partial f}{\partial x}(x_i, \theta) \right]^{-1} \frac{\partial f}{\partial \theta}(x_i, \theta). \quad (\text{A.6})$$

Choose $M > \sup U_i$ for every i ; if the second hypothesis holds, ensure also that M is large enough to have

$$|f(x, \theta_0)| > \frac{m}{2} \quad \forall |x| > M \quad \text{and} \quad \mathbb{E}[L(X)^{p'} I_{|X|>M}] < \varepsilon. \quad (\text{A.7})$$

On the compact set $K = [-M, M] \setminus \cup_i U_i$, by construction $f(\cdot, \theta_0)$ has a positive infimum; by uniform continuity of f on compact sets, we deduce that if $\delta'' > 0$ is sufficiently small, then for $(x, \theta) \in [-M, M] \times (\theta_0 - \delta'', \theta_0 + \delta'')$, f equals zero if and only if $x = \xi_i(\theta)$ for some i . In particular, if we define for ease of notation

$$\xi_0(\theta) = -M, \quad \xi_{k+1}(\theta) = M, \quad (\text{A.8})$$

then exactly one of the following holds:

- (a) $\{f(\cdot, \theta) > 0\} \cap (-M, M) = (\xi_0(\theta), \xi_1(\theta)) \cup \dots \cup (\xi_{2\lfloor \frac{k}{2} \rfloor}(\theta), \xi_{2\lfloor \frac{k}{2} \rfloor + 1}(\theta))$;
- (b) $\{f(\cdot, \theta) > 0\} \cap (-M, M) = (\xi_1(\theta), \xi_2(\theta)) \cup \dots \cup (\xi_{2\lfloor \frac{k+1}{2} \rfloor - 1}(\theta), \xi_{2\lfloor \frac{k+1}{2} \rfloor}(\theta))$.

As a consequence, in

$$\mathbb{E} [g(X) I_{f(X,\theta)>0} I_{|X|<M}] = \sum_{i=0}^k \int_{\xi_i(\theta)}^{\xi_{i+1}(\theta)} I_{f(x,\theta)>0} g(x) \rho(x) dx \quad (\text{A.9})$$

only one integral every two is non-trivial, and for that integral the indicator in the integrand is redundant and can be removed. Each $\xi_i(\theta)$ for $i = 1, \dots, k$ appears in exactly one of these nontrivial integrals, as left integration boundary if $(\partial f / \partial x)(\theta, \xi_i(\theta)) > 0$ and as right integration boundary otherwise. Hence, the fundamental theorem of calculus yields

$$\begin{aligned} \frac{\partial}{\partial \theta} \Big|_{\theta=\theta_0} \mathbb{E} [g(X) I_{f(X,\theta)>0} I_{|X|<M}] &= - \sum_{i=1}^k \text{sign} \left(\frac{\partial f}{\partial x}(\xi_i(\theta_0), \theta_0) \right) \xi_i'(\theta_0) g(\xi_i(\theta_0)) \rho(\xi_i(\theta_0)) \\ &= \sum_{i=1}^k \left| \frac{\partial f}{\partial x}(x_i, \theta_0) \right|^{-1} \frac{\partial f}{\partial \theta}(x_i, \theta_0) g(x_i) \rho(x_i). \end{aligned} \quad (\text{A.10})$$

To conclude we must prove that the integral on $\mathbb{R} \setminus (-M, M)$ does not contribute to the derivative. Under hypothesis 1 this is obvious. Otherwise, by definition of m, M and using the inequalities of Hölder and Markov,

$$\begin{aligned} & \left| \mathbb{E} [g(X) (I_{f(X,\theta_0+\Delta\theta)>0} - I_{f(X,\theta_0)>0}) I_{|X|>M}] \right| \\ & \leq \mathbb{E} |g(X) I_{|f(X,\theta_0+\Delta\theta)-f(X,\theta_0)|>m/2} I_{|X|>M}| \leq \mathbb{E} |g(X) I_{L(X)>(2\Delta\theta)^{-1}m} I_{|X|>M}| \\ & \leq \|g\|_p \mathbb{P} (L(X) > (2\Delta\theta)^{-1}m, |X| > M)^{1/p'} \leq \|g\|_p \left(\varepsilon m^{-p'} (2\Delta\theta)^{p'} \right)^{1/p'} =: C \varepsilon^{1/p'} \Delta\theta, \end{aligned} \quad (\text{A.11})$$

therefore

$$\liminf_{\Delta\theta \rightarrow 0} \Delta\theta^{-1} \mathbb{E} [g(X) (I_{f(X,\theta_0+\Delta\theta)>0} - I_{f(X,\theta_0)>0})] \geq RHS_{thesis} - C \varepsilon^{1/p'}, \quad (\text{A.12})$$

$$\limsup_{\Delta\theta \rightarrow 0} \Delta\theta^{-1} \mathbb{E} [g(X) (I_{f(X,\theta_0+\Delta\theta)>0} - I_{f(X,\theta_0)>0})] \leq RHS_{thesis} + C \varepsilon^{1/p'}, \quad (\text{A.13})$$

which by arbitrariness of ε concludes the proof. $\square$

Remark A.2. For sure one can imagine different conditions of integrability and asymptotic behaviour of f and g to guarantee that zeros appearing from infinity are irrelevant at first order; however, some condition is needed, as the following example shows.

Example A.3. Let $X \sim \mathcal{N}(0, 1)$, and denote by $\mathcal{N}$ its cumulative distribution function. Define the function $\phi \in C^1(\mathbb{R}^+)$ in this way:

$$\phi(r) = \begin{cases} \left(\frac{3}{2} - \frac{r^2}{2} \right) \exp \left(-\frac{r^2}{2} \right) & r \leq 1, \\ r^{-1} \exp \left(-\frac{r^2}{2} \right) & r > 1; \end{cases} \quad (\text{A.14})$$

note that for every $\theta > 0$ sufficiently small, there exists a unique counter-image $\phi^{-1}(\theta)$, which tends to $+\infty$ for $\theta \rightarrow 0$.

The functions $g(x) = 1$ and $f(x, \theta) = \phi(|x|) - \theta$ for $\theta_0 = 0$ satisfy all the hypotheses of the Lemma (with $k = 0$) except for hypotheses 1. and 2.; here we show that in this

case $\mathbb{E} [g(X)I_{f(X,\theta)>0}]$ is not differentiable in θ_0 . The left derivative is obviously zero. Instead, the right derivative is:

$$\lim_{\theta \downarrow 0} \frac{1}{\theta} \mathbb{E} [I_{\phi(|x|)-\theta>0}] = \lim_{\theta \downarrow 0} \frac{2}{\theta} [1 - \mathcal{N}(\phi^{-1}(\theta))]; \quad (\text{A.15})$$

to conclude that this derivative has a (finite and) non-zero value, we can use well known bounds on the tail of the Gaussian distribution function (see e.g. Duembgen, 2010), according to which $2\theta^{-1} [1 - \mathcal{N}(\phi^{-1}(\theta))]$ belongs to the following interval:

$$\frac{2}{\theta} \exp\left(-\frac{\phi^{-1}(\theta)^2}{2}\right) \frac{1}{\sqrt{2\pi}\phi^{-1}(\theta)} \left[1 - \frac{1}{\phi^{-1}(\theta)}, 1\right] \rightarrow \left\{\sqrt{\frac{2}{\pi}}\right\}. \quad (\text{A.16})$$

Equations like (3.6) follow immediately from the unidimensional Lemma, provided that it can be applied under the integral sign. In concrete cases this is almost always a routine verification, hence we will not pursue the maximum generality. We will limit ourselves to a couple of simple corollaries, which in many cases are more than enough. To this aim, we introduce the following hypothesis:

Hypothesis A.4.

1. $\mathbf{Z}$ is an $\mathbb{R}^d$ -valued random vector, and its k -th component has a conditional density $\rho(z_k|\mathbf{Z}_{-k})$ which is continuous in z_k for almost-every $\mathbf{Z}_{-k}$.
2. f is a function of variables $(\mathbf{z}, \theta)$, g is a function such that $g(\mathbf{Z}) \in L^1$.
3. For an appropriate $\varepsilon > 0$, almost every $\mathbf{Z}_{-k}$, and $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, the function

$$(\theta, z_k) \mapsto f(\theta, \mathbf{Z}(Z_k = z_k)) \quad (\text{A.17})$$

is C^1 , and for θ fixed it is equal to zero only for z_k in a finite set $A_k(\theta, \mathbf{Z}_{-k})$; on this set, the derivative with respect to z_k is non-null and the function $z_k \mapsto g(\mathbf{Z}(Z_k = z_k))$ is continuous.

4. Defining the function $D(\theta, \mathbf{Z}_{-k})$ as follows:

$$\sum_{\zeta_k \in A_k(\theta, \mathbf{Z}_{-k})} \rho(\zeta_k|\mathbf{Z}_{-k}) \left[\left| \frac{\partial f}{\partial z_k} \right|^{-1} \frac{\partial f}{\partial \theta} \right] (\theta, \mathbf{Z}(Z_k = \zeta_k)) \cdot g(\mathbf{Z}(Z_k = \zeta_k)), \quad (\text{A.18})$$

there exists $A \in L^1$ such that

$$|D(\theta, \mathbf{Z}_{-k})| \leq A \quad \forall \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon) \quad \text{for a.e. } \mathbf{Z}_{-k}. \quad (\text{A.19})$$

(Note the strict similarity with Hypothesis 3.1. However, here we add some technical condition to ensure that we can justify all the computations and apply the Lemma. On the other hand, $\mathbf{Z}$ here can be non-Gaussian.)

Corollary A.5. *Under the Hypothesis A.4, suppose that for almost-every $\mathbf{Z}_{-k}$ the cardinality of $A_k(\theta, \mathbf{Z}_{-k})$ is constant as θ varies in $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$. Then*

$$\frac{\partial}{\partial \theta} \mathbb{E} [I_{f(\theta, \mathbf{Z})>0} \cdot g(\mathbf{Z})] = \mathbb{E} [D(\theta, \mathbf{Z}_{-k})]. \quad (\text{A.20})$$

Corollary A.6. *Under the Hypothesis A.4, suppose that*

$$\liminf_{z \rightarrow \infty} |f(\theta_0, \mathbf{Z}(Z_k = \zeta))| > 0 \quad \text{for almost-every } \mathbf{Z}_{-k},$$

and that there exists a $p > 1$ such that the following integrability conditions hold:

- $g(\mathbf{Z}) \in L^p$;
- $|f(\theta_1, \mathbf{z}) - f(\theta_2, \mathbf{z})| \leq L(\mathbf{z})|\theta_1 - \theta_2|$ for every $\mathbf{z}$ and every $\theta_1, \theta_2 \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$, where L is a function such that $L(\mathbf{Z}) \in L^{p'}$.

Then

$$\frac{\partial}{\partial \theta} \mathbb{E} [I_{f(\theta, \mathbf{Z}) > 0} \cdot g(\mathbf{Z})] = \mathbb{E} [D(\theta, \mathbf{Z}_{-k})]. \quad (\text{A.21})$$

Proof. The tower rule gives

$$\mathbb{E} [I_{f(\theta, \mathbf{Z}) > 0} \cdot g(\mathbf{Z})] = \mathbb{E} [\mathbb{E} [I_{f(\theta, \mathbf{Z}) > 0} \cdot g(\mathbf{Z}) | \mathbf{Z}_{-k}]]. \quad (\text{A.22})$$

In both corollaries, the first three points in Hypothesis A.4 jointly with the hypothesis specific of the corollary are exactly what is needed to apply Lemma A.1 to the inner expectation when expressed using the conditional density in the Hypothesis. (The integrability conditions are valid under the conditional law for almost-every $\mathbf{Z}_{-k}$ thanks to the tower rule.) So, for a suitable version of that conditional expectation,

$$\frac{\partial}{\partial \theta} \mathbb{E} [I_{f(\theta, \mathbf{Z}) > 0} \cdot g(\mathbf{Z}) | \mathbf{Z}_{-k}] = D(\theta, \mathbf{Z}_{-k}) \quad \forall \theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon), \text{ for a.e. } \mathbf{Z}_{-k}. \quad (\text{A.23})$$

Now the last hypothesis allows the differentiation of (A.22) with respect to θ under the (outer) expectation sign by dominated convergence, which gives the conclusion. $\square$

With the following final examples we show how the two corollaries are used in concrete applications.

Example A.7 (Application of Corollary A.5). Let A_1, A_2 be independent assets with diffusive dynamics

$$dA_i(t) = rA_i(t)dt + \sigma_i(t, A_i(t))dW_i(t), \quad (\text{A.24})$$

where r is the (deterministic) risk-free rate, (W_1, W_2) is a standard bidimensional Wiener process, and $\sigma_i > 0$ are C^1 functions, equi-Lipschitz in their spatial argument. This guarantees in particular that the SDEs' solutions exist unique.

Let us consider the digital call on a basket, priced with Euler simulation (the limitation to two uncorrelated underlyings is only for notational simplicity):

$$P = \mathbb{E}[I_{A_1(T)+A_2(T) > 2K}] = \mathbb{E} [\mathbb{E}_{T-\delta} [I_{A_1(T)+A_2(T) > 2K}]], \quad (\text{A.25})$$

where in the inner expected value

$$A_i(T) = A_i(T, Z_i) = (1 + r\delta)A_i(T - \delta) + \sigma_i(t, A_i(T - \delta))\sqrt{\delta}Z_i, \quad (\text{A.26})$$

in which $A_i(T - \delta)$ are known and Z_i are independent and identically distributed with law $\mathcal{N}(0, 1)$.

We would like to apply Corollary A.5 to compute the derivative with respect to $A_1(T - \delta)$ of the inner expected value in (A.25); the only non trivially true hypothesis is the last one in A.4. To show that it holds, we define $t = T - \delta$ and compute $D(A_1(t), Z_2)$:

$$D(A_1(t), Z_2) = \left(1 + r\delta + \frac{\partial \sigma_1(t, A_1(t))}{\partial A_1(t)} \frac{2K - A_2(T, Z_2) - (1 + r\delta)A_1(t)}{\sigma_1(t, A_1(t))} \right) \times \frac{\exp \left\{ -\frac{(2K - A_2(T, Z_2) - (1 + r\delta)A_1(t))^2}{2\delta \sigma_1(t, A_1(t))^2} \right\}}{\sqrt{2\pi\delta} \sigma_1(t, A_1(t))}, \quad (\text{A.27})$$

whose absolute value can be uniformly controlled in a neighbourhood of $A_1(t)$ with an integrable quantity, because by the Lipschitz property $\sigma_1(t, \cdot) > \varepsilon$ in a neighbourhood of $A_1(t)$ and therefore:

$$|D(A_1(t), Z_2)| \leq \left(1 + r\delta + \frac{L(C_1 + C_2|A_2(t)| + C_3|Z_2|)}{\varepsilon}\right) (\sqrt{2\pi\delta\varepsilon})^{-1} \quad (\text{A.28})$$

for suitable constants C_i .

Finally, the following identity is straightforwardly verified:

$$\frac{\partial}{\partial A_1(0)} \mathbb{E} [\mathbb{E}_{T-\delta} [I_{A_1(T)+A_2(T)>2K}]] = \mathbb{E} \left[\frac{\partial}{\partial A_1(0)} \mathbb{E}_{T-\delta} [I_{A_1(T)+A_2(T)>2K}] \right], \quad (\text{A.29})$$

so that at least in this case, we can conclude that the algorithm of Section 3.5 is completely justified.

Example A.8 (Application of Corollary A.6). With $\theta_0 > 0$, let $\mathbf{Z}$ be a standard bidimensional Gaussian vector, $g = 1$ and

$$f(\theta, \mathbf{z}) = \theta - e^{z_1} - e^{z_2}. \quad (\text{A.30})$$

This way, except for constants, $\mathbb{E} [I_{f(\theta, \mathbf{Z})>0} \cdot g(\mathbf{Z})]$ may be the price of a digital call on the average of two log-normal assets.

No $\varepsilon > 0$ exists such that the cardinality of $A_1(\theta, Z_2)$ is constant for $\theta \in (\theta_0 - \varepsilon, \theta_0 + \varepsilon)$ for almost every Z_2 : this is because for $Z_2 \in (\ln(\theta_0 - \varepsilon), \ln(\theta_0 + \varepsilon))$, the cardinality of that set is zero near $\theta = \theta_0 - \varepsilon$, while it is one near $\theta = \theta_0 + \varepsilon$. Therefore Corollary A.5 cannot be applied. However,

$$D(\theta, z_2) = \begin{cases} (\theta - e^{z_2})^{-1} \exp \left\{ -(\ln(\theta - e^{z_2}))^2 / 2 \right\} & \text{if } z_2 < \ln(\theta), \\ 0 & \text{otherwise.} \end{cases} \quad (\text{A.31})$$

Elementary estimates show that $|D(\theta, Z_2)|$ is lower than a suitable L^1 random variable uniformly for each θ in a neighbourhood of θ_0 ; all the other hypotheses of Corollary A.6 are trivially satisfied.

A.3 Call spread analytical sensitivities

We provide here for the sake of completeness a derivation of the explicit formulas used in Section 3.6.1 for the variance of the sample estimators for the Delta and Vega sensitivities of a call spread. We will use the following notations:

$$\tilde{A}_{tT} := A_t \exp \left\{ \int_t^T \mu(u) du \right\}, \quad \Sigma_{tT} := \sqrt{\int_t^T \sigma^2(u) du}. \quad (\text{A.32})$$

A.3.1 Delta sensitivity

The pathwise estimator for the Delta sensitivity of the call spread is given by

$$\delta CS_t^{(N)}(A_t, K, \varepsilon) := \frac{P(t, T)}{N} \sum_{i=1}^N \frac{A_T}{A_t} \frac{H(A_T(\omega_i) - K_-) - H(A_T(\omega_i) - K_+)}{\varepsilon}. \quad (\text{A.33})$$

Since $CS_t^{(N)}$ is an average of N independent and identically distributed samples, we have that

$$\text{Var} \left[CS_t^{(N)} \right] = N^{-1} \text{Var} \left[CS_t^{(1)} \right]. \quad (\text{A.34})$$

To compute the right hand side, we expand the following square:

$$\delta CS_t^{(1)}(A_t, K, \varepsilon)^2 = P(t, T)^2 \frac{A_T^2}{A_t^2} \frac{H(A_T - K_-)^+ - H(A_T - K_+)^+}{\varepsilon^2}; \quad (\text{A.35})$$

the key quantities to compute have the form of

$$\mathbb{E} \left[A_T^2 H(A_T - K)^+ \right] = \mathbb{E} \left[A_T^2 H(A_T^2 - K^2)^+ \right] \quad (\text{A.36})$$

which is the undiscounted price of a cash-or-nothing call with strike K^2 on the log-normal asset A_T^2 , whose instantaneous volatility is $2\sigma(s)$ and whose T -forward value is

$$F_{tT}^{(2)} = \tilde{A}_{tT}^2 e^{\Sigma_{tT}^2}. \quad (\text{A.37})$$

Hence (see e.g. Duffie, 2001)

$$\mathbb{E} \left[A_T^2 H(A_T - K)^+ \right] = F_{tT}^{(2)} \mathcal{N} \left(d_+ \left(K^2, F_{tT}^{(2)}, 2\Sigma_{tT} \right) \right), \quad (\text{A.38})$$

and finally the second moment of our estimator is

$$\begin{aligned} \mathbb{E} \left[\delta CS_t^{(1)}(A_t, K, \varepsilon)^2 \right] &= \\ P(t, T)^2 \frac{F_{tT}^{(2)}}{A_t^2} \varepsilon^{-2} &\left[\mathcal{N} \left(d_+ \left(K_-^2, F_{tT}^{(2)}, 2\Sigma_{tT} \right) \right) - \mathcal{N} \left(d_+ \left(K_+^2, F_{tT}^{(2)}, 2\Sigma_{tT} \right) \right) \right], \end{aligned} \quad (\text{A.39})$$

which is $O(\varepsilon^{-1})$ since the difference in squared brackets is a finite difference of a smooth function between two points K_-^2 and K_+^2 , hence it is $O(K_-^2 - K_+^2) = O(\varepsilon)$.

Eventually, noting that $\mathbb{E}[\delta CS_t^{(1)}] = \delta CS_t + O(1)$, we can conclude that also the variance

$$\begin{aligned} \text{Var} \left[CS_t^{(1)} \right] &= \mathbb{E} \left[\delta CS_t^{(1)}(A_t, K, \varepsilon)^2 \right] - \mathbb{E} \left[\delta CS_t^{(1)}(A_t, K, \varepsilon) \right]^2 \\ &= P(t, T)^2 \frac{F_{tT}^{(2)}}{A_t^2} \varepsilon^{-2} \left[\mathcal{N} \left(d_+ \left(K_-^2, F_{tT}^{(2)}, 2\Sigma_{tT} \right) \right) - \mathcal{N} \left(d_+ \left(K_+^2, F_{tT}^{(2)}, 2\Sigma_{tT} \right) \right) \right] \\ &\quad - P(t, T)^2 \frac{\tilde{A}_{tT}^2}{A_t^2} \varepsilon^{-2} \left[\mathcal{N} \left(d_+ \left(K_-, \tilde{A}_{tT}, \Sigma_{tT} \right) \right) - \mathcal{N} \left(d_+ \left(K_+, \tilde{A}_{tT}, \Sigma_{tT} \right) \right) \right]^2 \end{aligned} \quad (\text{A.40})$$

is $O(\varepsilon^{-1})$ as desired.

A.3.2 Vega sensitivity

We perform the same kind of computation for the Vega sensitivity pathwise estimator

$$\begin{aligned} \nu CS_t^{(N)}(A_t, K, \varepsilon) &:= \\ \frac{P(t, T)}{N} \sum_{i=1}^N A_T(\omega_i) [-\sigma(T-t) + (W_T - W_t)] &\frac{H(A_T(\omega_i) - K_-) - H(A_T(\omega_i) - K_+)}{\varepsilon}. \end{aligned} \quad (\text{A.41})$$

The square of $\nu CS_t^{(1)}$ is

$$\frac{P(t, T)^2 A_{tT}^2}{\varepsilon^2} [-\sigma(T - t) + (W_T - W_t)]^2 \cdot [H(A_T - K_-) - H(A_T - K_+)], \quad (\text{A.42})$$

whose expected value is

$$\frac{P(t, T)^2 \tilde{A}_{tT}^2}{\varepsilon^2} [I(K_-, T - t, \sigma) - I(K_+, T - t, \sigma)] \quad (\text{A.43})$$

with the definition

$$I(K, \tau, \sigma) := \int_k^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{-\sigma^2 \tau} e^{2\sigma \sqrt{\tau} x} [-\sigma \tau + \sqrt{\tau} x]^2 dx, \quad (\text{A.44})$$

where

$$k = \frac{1}{\sigma \sqrt{\tau}} \log \left(\frac{K}{\tilde{A}_{tT}} \right) + \frac{1}{2} \sigma \sqrt{\tau} = -d_-(K). \quad (\text{A.45})$$

To compute this integral, we make the change of variables $y = x - 2\sigma \sqrt{\tau}$:

$$I(K, \tau, \sigma) = e^{\sigma^2 \tau} \tau \int_{k-2\sigma \sqrt{\tau}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} [\sigma \sqrt{\tau} + y]^2 dy. \quad (\text{A.46})$$

Now we put $a := k - 2\sigma \sqrt{\tau}$ and use the three elementary integrals

$$\begin{aligned} \int_a^\infty e^{-y^2/2} dy &= \sqrt{2\pi} (1 - \mathcal{N}(a)), \\ \int_a^\infty y e^{-y^2/2} dy &= e^{-a^2/2}, \\ \int_a^\infty y^2 e^{-y^2/2} dy &= -y e^{-y^2/2} \Big|_a^\infty + \int_a^\infty e^{-y^2/2} dy = a e^{-a^2/2} + \sqrt{2\pi} (1 - \mathcal{N}(a)), \end{aligned} \quad (\text{A.47})$$

where $\mathcal{N}$ is the standard normal cumulative distribution function, getting

$$\begin{aligned} I(K, \tau, \sigma) &= e^{\sigma^2 \tau} \tau \left\{ (1 - \mathcal{N}(a)) (\sigma^2 \tau + 1) + \frac{1}{\sqrt{2\pi}} e^{-a^2/2} (a + 2\sigma \sqrt{\tau}) \right\} \\ &= e^{\sigma^2 \tau} \tau \left\{ (1 - \mathcal{N}(k - 2\sigma \sqrt{\tau})) (\sigma^2 \tau + 1) + \frac{k}{\sqrt{2\pi}} e^{-(k-2\sigma \sqrt{\tau})^2/2} \right\}. \end{aligned} \quad (\text{A.48})$$

We plug this into (A.43) to get

$$\begin{aligned} \mathbb{E} [\nu CS_t^{(1)}(A_t, K, \varepsilon)^2] &= \frac{P(t, T)^2 \tilde{A}_{tT}^2 e^{\sigma^2(T-t)} (T-t)}{\varepsilon^2} \times \\ &\quad \left\{ \left(\mathcal{N}(k_+ - 2\sigma \sqrt{T-t}) - \mathcal{N}(k_- - 2\sigma \sqrt{T-t}) \right) (\sigma^2(T-t) + 1) + \right. \\ &\quad \left. \frac{k_-}{\sqrt{2\pi}} e^{-\frac{(k_- - 2\sigma \sqrt{T-t})^2}{2}} - \frac{k_+}{\sqrt{2\pi}} e^{-\frac{(k_+ - 2\sigma \sqrt{T-t})^2}{2}} \right\}, \end{aligned} \quad (\text{A.49})$$

which is $O(\varepsilon^{-1})$ since the difference in curly brackets is a finite difference of a smooth function between two points K_- and K_+ with distance ε .

Eventually, the variance is readily computed as

$$\text{Var} [\nu CS_t^{(1)}(A_t, K, \varepsilon)] = \mathbb{E} [\nu CS_t^{(1)}(A_t, K, \varepsilon)^2] - \mathbb{E} [\nu CS_t^{(1)}(A_t, K, \varepsilon)]^2, \quad (\text{A.50})$$

in which by construction

$$\mathbb{E} \left[\nu C S_t^{(1)}(A_t, K, \varepsilon) \right] = \nu C S_t = \varepsilon^{-1} A_t \sqrt{\frac{T-t}{2\pi}} \left[e^{-d_+(K_-)^2/2} - e^{-d_+(K_+)^2/2} \right] = O(1), \quad (\text{A.51})$$

where we used the well known formula for the Vega sensitivity of a vanilla call.

A.4 Second order distributional differentiation

In this appendix, we will derive an estimator for the Hessian of the price

$$P = \mathbb{E} [g(\mathbf{X}_T, \boldsymbol{\theta})]$$

for a piecewise smooth but possibly discontinuous payoff g , by a generalization of the Dirac-delta based method DAAD of Chapter 3.

In view of (4.9), we just need to differentiate twice the conditional expectation ε of Section 4.4.2, which has the following form:

$$\varepsilon = \mathbb{E} [\phi(I_{f_1(\boldsymbol{\theta}, \mathbf{Z}) > 0}, \dots, I_{f_h(\boldsymbol{\theta}, \mathbf{Z}) > 0}, \boldsymbol{\theta}, \mathbf{Z})] = \mathbb{E} [\phi(\mathbf{I}_{\mathbf{f}(\boldsymbol{\theta}, \mathbf{Z}) > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z})],$$

where $\mathbf{Z}$ is a standard Gaussian random vector, and ϕ and the f_i are suitable functions, smooth in $\boldsymbol{\theta}$ and $\mathbf{Z}$. We make the following weak assumption:

Hypothesis A.9. f_i is twice differentiable with respect to $\boldsymbol{\theta}$ and $z_{k(i)}$. Moreover, for almost every $\mathbf{Z}_{-k(i)}$, the function $z_{k(i)} \mapsto f(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = z_{k(i)}))$ is zero only for $z_{k(i)}$ in a finite set $A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})$, and its derivative in such points is non-null.

We also suppose for notational convenience and with no loss of generality that

$$\Delta^{(i)} \phi(\mathbf{I}_{\mathbf{f}(\boldsymbol{\theta}, \mathbf{Z}) > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z}) = 0 \quad \forall i < c,$$

so that the indicator functions before c represent the discontinuities in the gradient of the payoff, while the indicator functions from c onwards represent the discontinuities of the payoff itself, if any.

Equation (3.11) tells us that the first order sensitivity is

$$\frac{\partial \varepsilon}{\partial \boldsymbol{\theta}} = \mathbf{d}_0 + \sum_{i=c}^h \mathbf{d}_i,$$

where

$$\begin{aligned} \mathbf{d}_0 &= \mathbb{E} \left[\frac{\partial \phi}{\partial \boldsymbol{\theta}}(\mathbf{I}_{\mathbf{f}(\boldsymbol{\theta}, \mathbf{Z}) > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z}) \right], \\ \mathbf{d}_i &= \mathbb{E} \left[\sum_{\zeta \in A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})} \left\{ \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \right. \right. \\ &\quad \left. \left. \times \Delta^{(i)} \phi(\mathbf{I}_{\mathbf{f}(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \right\} \right]. \end{aligned}$$

Now we have to differentiate $\mathbf{d}_0$ and $\mathbf{d}_i$. The first task is a direct application of the first order result to $\nabla\phi$:

$$\frac{\partial \mathbf{d}_0}{\partial \boldsymbol{\theta}} = \mathbf{H}_{00} + \sum_{i=1}^h \mathbf{H}_{0i},$$

where

$$\begin{aligned} \mathbf{H}_{00} &= \mathbb{E} \left[\frac{\partial^2 \phi}{\partial \boldsymbol{\theta}^2} \left(\mathbf{I}_{\mathbf{f}(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)}=\zeta)) > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z} \right) \right], \\ \mathbf{H}_{0i} &= \mathbb{E} \left[\sum_{\zeta \in A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})} \left\{ \Delta^{(i)} \left(\frac{\partial \phi}{\partial \boldsymbol{\theta}} \right)^\top \left(\mathbf{I}_{\mathbf{f}(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)}=\zeta)) > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)}=\zeta) \right) \right. \right. \\ &\quad \left. \left. \times \frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left[\left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)}=\zeta)) \right\} \right]. \end{aligned}$$

If the payoff is continuous ($c = h + 1$), we are done; otherwise, we need the derivative of $\mathbf{d}_i$, which is slightly trickier. In order to compute it, we note that by the implicit function theorem, in a neighbourhood of $\boldsymbol{\theta}$, the zeros ζ in the definition of each $\mathbf{d}_i$ can be expressed as functions $\zeta_i^{(l)}$ of $\mathbf{Z}_{-k(i)}$ and $\boldsymbol{\theta}$ with gradient¹

$$- \left[\left(\frac{\partial f_i}{\partial z_{k(i)}} \right)^{-1} \left(\frac{\partial f_i}{\partial (\boldsymbol{\theta}, \mathbf{z}_{-k(i)})} \right) \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)}=\zeta)).$$

Now each summand in the definition of $\mathbf{d}_i$ can be differentiated with the DAAD method, provided that the composition of f_j with $\zeta_i^{(l)}$ satisfies Hypothesis A.9 for some coordinate $h(i, j) \neq k(i)$. We define

$$A^{i,j}(\boldsymbol{\theta}, \mathbf{Z}) := \{(\zeta, \eta) : (f_i, f_j)(\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)}=\zeta, Z_{h(i,j)}=\eta)) = (0, 0)\}.$$

We are ready to state the result:

$$\frac{\partial \mathbf{d}_i}{\partial \boldsymbol{\theta}} = \mathbf{H}_{i0} + \sum_{j=c}^h \mathbf{H}_{ij},$$

¹In fact the neighbourhood may depend on $\mathbf{Z}$; we neglect this kind of technicalities, which should be addressable along the lines of Appendix A.2.

where

$$\begin{aligned} \mathbf{H}_{i0} &= \mathbb{E} \left\{ \sum_{\zeta \in A_{k(i)}^i(\boldsymbol{\theta}, \mathbf{Z}_{-k(i)})} \frac{\partial}{\partial(\boldsymbol{\theta}, z_{k(i)})} \Big|_{z_{k(i)}=\zeta} \left[\frac{e^{-\zeta^2/2}}{\sqrt{2\pi}} \left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \left(\frac{\partial f_i}{\partial \boldsymbol{\theta}} \right)^\top \Delta^{(i)} \phi \right] \right. \\ &\quad \times \left. \left(- \left[\left(\frac{\partial f_i}{\partial z_{k(i)}} \right)^{-1} \left(\frac{\partial f_i}{\partial \boldsymbol{\theta}} \right) \right] (\boldsymbol{\theta}, \mathbf{Z}(Z_{k(i)} = \zeta)) \right) \right\}, \\ \mathbf{H}_{ij} &= \mathbb{E} \left\{ \sum_{(\zeta, \eta) \in A^{i,j}(\boldsymbol{\theta}, \mathbf{Z})} \left[\frac{e^{-\frac{\zeta^2 + \eta^2}{2}}}{2\pi} \left| \frac{\partial f_i}{\partial z_{k(i)}} \right|^{-1} \left(\frac{\partial f_i}{\partial \boldsymbol{\theta}} \right)^\top \left| \frac{\partial f_j}{\partial z_{h(i,j)}} - \frac{\partial f_j}{\partial z_{k(i)}} \left(\frac{\partial f_i}{\partial z_{k(i)}} \right)^{-1} \frac{\partial f_i}{\partial z_{h(i,j)}} \right|^{-1} \right. \right. \\ &\quad \left. \left. \left(\frac{\partial f_j}{\partial \boldsymbol{\theta}} - \frac{\partial f_j}{\partial z_{k(i)}} \left(\frac{\partial f_i}{\partial z_{k(i)}} \right)^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \right) \Delta^{(j)} \Delta^{(i)} \phi(\mathbf{I}_{\mathbf{f} > \mathbf{0}}, \boldsymbol{\theta}, \mathbf{Z}) \right] \right\}_{Z_{k(i)}=\zeta, Z_{h(i,j)}=\eta}. \end{aligned}$$

Note that the expression in the second absolute value should be different from zero, or to say it differently, $\partial(f_i, f_j)/\partial(z_{k(i)}, z_{h(i,j)})$ should be full rank.

All the gradients appearing in these formulas are readily computed by algorithmic differentiation; the only practical concern is the fact that there are (up to) $(h - c + 1)^2$ terms $\mathbf{H}_{ij}$. In fact in most applications, f_i and f_j depend on non-overlapping sets of fixing times, so that

$$\begin{aligned} \left| \frac{\partial f_j}{\partial z_{h(i,j)}} - \frac{\partial f_j}{\partial z_{k(i)}} \left(\frac{\partial f_i}{\partial z_{k(i)}} \right)^{-1} \frac{\partial f_i}{\partial z_{h(i,j)}} \right|^{-1} \left(\frac{\partial f_j}{\partial \boldsymbol{\theta}} - \frac{\partial f_j}{\partial z_{k(i)}} \left(\frac{\partial f_i}{\partial z_{k(i)}} \right)^{-1} \frac{\partial f_i}{\partial \boldsymbol{\theta}} \right) \\ = \left| \frac{\partial f_j}{\partial z_{k(j)}} \right|^{-1} \left(\frac{\partial f_j}{\partial \boldsymbol{\theta}} \right), \end{aligned}$$

which does not depend on i any more. As a consequence, the only portion of the algorithm which truly grows quadratically with the number of payoff discontinuities is $\Delta^{(j)} \Delta^{(i)} \phi$. This should not be dramatic in the economy of a Monte Carlo pricing exercise inclusive of a costly Euler simulation, unless there is a really large number of digital features as in frequently monitored barrier options, which anyway are often already handled by regularizing the payoff accounting for touch probabilities (see e.g. Glasserman, 2004).

For ease of reference, we collect as a conclusion the full set of addends contributing to the estimator of the Hessian:

$$\frac{\partial^2 \varepsilon}{\partial \boldsymbol{\theta}^2} = \mathbf{H}_{00} + \sum_{i=1}^h \mathbf{H}_{0i} + \sum_{i=c}^h \mathbf{H}_{i0} + \sum_{i,j=c}^h \mathbf{H}_{ij}.$$

Bibliography

- Alexander, C. and Nogueira, L. M. (2004). Hedging with stochastic and local volatility. ICMA Center Discussion Papers in Finance n.11.
- Bartlett, B. (2006). Hedging under SABR model. *Wilmott Magazine*.
- Basel Committee on Banking Supervision (2013). Fundamental review of the trading book: a revised market risk framework.
- Basel Committee on Banking Supervision (2015). Review of the credit valuation adjustment risk framework.
- Benhamou, E. (2003). Optimal Malliavin weighting function for the computation of the Greeks. *Mathematical Finance*, 13(1):37–53.
- Björk, T. (2009). *Arbitrage Theory in Continuous Time*. Oxford University Press, Oxford, third edition.
- Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81(3):637–654.
- Capriotti, L. (2008). Reducing the variance of likelihood ratio Greeks in Monte Carlo. In Mason, S. J., Hill, R. R., Mnch, L., Rose, O., Jefferson, T., and Fowler, J. W., editors, *Proceedings of the 2008 Winter Simulation Conference*, pages 587–593.
- Capriotti, L. (2011). Fast Greeks by algorithmic differentiation. *Journal of Computational Finance*, 14(3):3–35.
- Capriotti, L. (2015). Likelihood ratio method and algorithmic differentiation: Fast second order Greeks. *Algorithmic Finance*, 4:81–87.
- Capriotti, L. and Giles, M. (2012). Adjoint Greeks made easy. *Risk Magazine*, 25(95):92–99.
- Capriotti, L., Lee, J., and Peacock, M. (2011). Real time counterparty credit risk in Monte Carlo. *Risk Magazine*, 24(86).
- Chan, J. H. and Joshi, M. (2015). Optimal limit methods for computing sensitivities of discontinuous integrals including triggerable derivative securities. *IEEE Transactions*, 47:978–997.
- Crépey, S. (2004). Delta hedging vega risk? *Quantitative Finance*, 4(5):559–579.
- Daluiso, R. and Facchinetti, G. (2018). Algorithmic differentiation for discontinuous payoffs. *International Journal of Theoretical and Applied Finance*, 21(4).

- Daluiso, R. and Morini, M. (2017). Hedging efficiently under correlation. *Quantitative Finance*, 17(10):1535–1547.
- Duembgen, L. (2010). Bounding standard Gaussian tail probabilities. arXiv:1012.2063.
- Duffie, D. (2001). *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton.
- El Karoui, N., Jeanblanc-Picqué, M., and Shreve, S. S. (1998). Robustness of the Black and Scholes formula. *Mathematical Finance*, 8(2):93–126.
- Feng, G. and Liu, G. (2016). Conditional Monte Carlo: A change-of-variables approach. arXiv:1603.06378v1.
- Fournié, E., Lasry, J.-M., Lebuchoux, J., and Lions, P.-L. (2001). Applications of Malliavin calculus to Monte Carlo methods in finance, II. *Finance and Stochastics*, 5:201–236.
- Fournié, E., Lasry, J.-M., Lebuchoux, J., Lions, P.-L., and Touzi, N. (1999). Applications of Malliavin calculus to Monte Carlo methods in finance. *Finance and Stochastics*, 3:391–412.
- Fu, M. and Hu, M. J.-Q. (1997). *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic Publishers, New York.
- Giles, M. (2007). Monte Carlo evaluation of sensitivities in computational finance. Technical Report NA07/12, Oxford University Computing Lab.
- Giles, M. (2009). Vibrato Monte Carlo sensitivities. In L’Ecuyer, P. and Owen, A. B., editors, *Monte Carlo and Quasi-Monte Carlo Methods 2008*, pages 369–382, Berlin. Springer.
- Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer, Berlin.
- Griewank, A. and Walther, A. (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Society for Industrial and Applied Mathematics, Philadelphia, second edition.
- Henrard, M. (2011). Adjoint algorithmic differentiation: Calibration and implicit function theorem. OpenGamma Quantitative Research n.1.
- Henrard, M. (2013). Algorithmic differentiation in finance: Root finding and least square calibration. OpenGamma Quantitative Research n.7.
- Herrmann, S. and Muhle-Karbe, J. (2017). Model uncertainty, recalibration, and the emergence of delta-vega hedging. Swiss Finance Institute Research Paper No.15-52.
- Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.
- Hodges, S. D. and Neuberger, A. (1989). Optimal replication of contingent claims under transaction costs. *The Review of Futures Markets*, 8(2):222–242.

- Hull, J. and White, A. (1990). Pricing interest-rate derivative securities. *The Review of Financial Studies*, 3(4):573–592.
- International Swaps and Derivatives Association (2013). Standard initial margin model for non-cleared derivatives.
- International Swaps and Derivatives Association (2016). ISDA SIMM methodology, version R1.0.
- Joshi, M. and Yang, C. (2011). Algorithmic Hessians and the fast computation of cross-gamma risk. *IIE Transactions*, 43:878–892.
- Joshi, M. and Zhu, D. (2016). Optimal partial proxy method for computing gammas of financial products with discontinuous and angular payoffs. *Applied Mathematical Finance*, 23(1):22–56.
- Karatzas, I. and Shreve, S. E. (1991). *Brownian Motion and Stochastic Calculus*. Springer, New York.
- Kloeden, P. E. and Platen, E. (1992). *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin. Corrected Third Printing 1999.
- Liu, G. and Hong, L. J. (2011). Kernel estimation of the Greeks for options with discontinuous payoffs. *Operations Research*, 59(1):96–108.
- Lyu, Y.-D. and Teng, H.-W. (2011). Unbiased and efficient Greeks of financial options. *Finance and Stochastics*, 15:141–181.
- Malliavin, P. and Thalmaier, A. (2006). *Stochastic Calculus of Variations in Mathematical Finance*. Springer, Berlin.
- Mercurio, F. (2010). Interest rates and the credit crunch: New formulas and market models. Bloomberg Portfolio Research Paper No. 2010-01-FRONTIERS.
- Mercurio, F. and Morini, M. (2009). A note on hedging with local and stochastic volatility models. In Mercurio, F., editor, *Modelling Interest Rates: Advances in Derivatives Pricing*. Risk Books.
- Morini, M. (2011). *Understanding and Managing Model Risk. A Practical Guide for Quants, Traders and Validators*. Wiley Finance.
- Pagès, G., Pironneau, O., and Sall, G. (2016). Vibrato and automatic differentiation for high order derivatives and sensitivities of financial options. arXiv:1606.06143v1.
- Piterbarg, V. (2003). A practitioners guide to pricing and hedging callable LIBOR exotics in forward LIBOR models. Available at SSRN: <https://ssrn.com/abstract=427084> or <http://dx.doi.org/10.2139/ssrn.427084>.
- Reghai, A., Kettani, O., and Messaoud, M. (2015). CVA with Greeks and AAD. *Risk Magazine*.
- Schwartz, L. (1966). *Théorie des Distributions*. Hermann, Paris.
- Whalley, E. and Wilmott, P. (1997). An asymptotic analysis of an optimal hedging model for option pricing with transaction costs. *Mathematical Finance*, 7(3):307–324.
- Zakamouline, V. I. (2005). Optimal hedging of options with transaction costs. *Wilmott Magazine*.

Acknowledgements

I am very grateful to professors Massimo Morini and Fabio Bellini, who believed in the idea of me attending a PhD programme without leaving my job, and made everything they could to help it become a reality. Massimo has been a real research mentor throughout these years, and while his signature formally appears only in the first chapter, his suggestions were very valuable throughout this project. Fabio's help and advice was also crucial as a guide in the world of the University of Milano-Bicocca and of academia in general.

I also want to thank the head of Banca IMI's Financial Engineering Andrea Bugin for supporting my initiative.

Finally, many colleagues would deserve a share of the merit of this thesis because of useful discussions on its contents: thanks to Andrea Prampolini, Michele Trapletti, Marco Pincioli, Francesco Mandelli, Andrea Pallavicini, Nicola Moreni, Giuseppe Di Poto, Giulio Sartorelli... but if I had to mention only one, no doubt I am indebted the most to Giorgio Facchinetti: not only for his scientific and artistic contributions to Chapter 3 (and elsewhere); but perhaps even more, because he understands very well that time dedicated to research is not wasted time, and bears with a collaborator who sometimes disappears from his desk to go to a seminar or a conference.

Milan, December 29, 2018,
Roberto Daluiso